2015-1200

IN THE

UNITED STATES COURT OF APPEALS

FOR THE FEDERAL CIRCUIT

INDUSTRIAL TECHNOLOGY RESEARCH INSTITUTE,
Appellant,

v.

PACIFIC BIOSCIENCES OF CALIFORNIA, INC.,
Appellee.

Appeal from the United States Patent and Trademark Office Patent Trial and Appeal Board in Patent Interference No. 105,970, Administrative Patent Judges John G. New, Sally Gardner Lane, and James T. Moore

---

BRIEF OF APPELLANT INDUSTRIAL TECHNOLOGY RESEARCH INSTITUTE

David S. Forman
FINNEGAN, HENDERSON, FARABOW, GARRETT & DUNNER, LLP
901 New York Avenue, NW
Washington, DC 20005-4413
(202) 408-4000

Erik R. Puknys
Adam M. Breier
FINNEGAN, HENDERSON, FARABOW, GARRETT & DUNNER, LLP
3300 Hillview Avenue
Palo Alto, CA 94304-1203
(650) 849-6600

Attorneys for Appellants

March 19, 2015

CERTIFICATE OF INTEREST

Counsel for Appellant Industrial Technology Research Institute certifies the following:

1. The full name of every party or amicus represented by me is:

Industrial Technology Research Institute

Ti-Shiue Biotech, Inc.

2. The name of the real party in interest represented by me is:

None

3. All parent corporations and any publicly held companies that own 10 percent or more of the stock of the party represented by me are:

Personal Genomics, Inc.

4. The names of all law firms and the partners or associates that appeared for the party or amicus now represented by me in the trial court or agency or are expected to appear in this court are:

FINNEGAN, HENDERSON, FARABOW, GARRETT & DUNNER, LLP
David S. Forman
Erik R. Puknys
M. Paul Barker
Adam M. Breier

TABLE OF CONTENTS

TABLE OF AUTHORITIES

CASES

STATUTES

RULES

STATEMENT OF RELATED CASES

No other appeal in or from this action was previously before this or any other appellate court. Counsel knows of no other case pending in this Court or any other court that may directly affect, or be directly affected by, the Court's decision in this appeal.

I. JURISDICTION

The Patent Trial and Appeal Board (Board) had jurisdiction under 35 U.S.C. § 135(a). The Board rendered its final judgment on September 3, 2014, in favor of Pacific Biosciences of California, Inc. (PacBio). Industrial Technology Research Institute and Ti-Shiue Biotech, Inc. (collectively, ITRI) timely appealed on October 31, 2014. This Court has jurisdiction pursuant to 28 U.S.C. § 1295(a)(4)(A) and 35 U.S.C. § 141.

II. STATEMENT OF THE ISSUES

1. Whether the Board erred in concluding that PacBio was entitled to the benefit of Application No. 61/201,551, where the ’551 application lacks a written description of the Count.

2. Whether the Board erred in concluding that claims 1-26 of the ’630 patent would have been obvious by analyzing the prior art with improper hindsight. Whether the Board additionally erred in concluding that claim 23, which requires accepting or rejecting a sample sequence based on scores of a different sequence, would have been obvious over prior art that did not teach that step.

3. Whether the Board erred in concluding that claims 27 and 28 of the ’630 patent, which require a nucleotide analog that discriminates between a base and its modified form, would have been obvious over a combination of references

that did not disclose a discriminating nucleotide analog and where the prior art taught away from the claimed invention.

4. Whether the Board erred in failing to hold that the prior art asserted by PacBio rendered PacBio's own claims unpatentable under the presumption of cross-applicability, 37 C.F.R. § 41.207(c).

III. PRELIMINARY STATEMENT

From its founding, this Court has warned against the dangers of hindsight reasoning. This appeal illustrates those dangers and shows that even today, the agency entrusted to grant patents can fall prey to hindsight's temptations.

Here, ITRI invented a new way to locate modified bases in a DNA molecule. Modified bases differ from ordinary DNA molecule bases (adenine, cytosine, guanine, and thymine) and, because of their effects on hereditary diseases, are the subject of intense scientific interest. But to study these modified bases, researchers have to locate them. Before ITRI's invention, the most common way to locate modified bases required a reference DNA molecule to have been sequenced (the constituent bases and their locations identified) first. Then the molecule would be subjected to a chemical reaction and sequenced again. The sequence of the chemically altered molecule would then be compared to the reference sequence and, based on whether certain bases were resistant or susceptible to chemical alteration, the modified bases could be located.

This prior-art method had limits, however. Specifically, it required reference DNA, so researchers either had to study DNA molecules that had already been sequenced or determine the reference sequence themselves. Particularly when single molecule sequencing was used, sequencing was preferably performed multiple times—in a process called "consensus sequencing"—before sequence data could be considered sufficiently reliable. Consensus sequencing often used sequences of both the forward and reverse strands of the DNA double helix because bases at corresponding positions in these strands normally match under the Watson-Crick rules (adenine matches with thymine, and cytosine matches with guanine). Matches in forward and reverse strands indicated that the sequencing was accurate (reliable), and mismatches indicated sequencing errors.

ITRI created an improved method for locating modified bases that did not require a reference sequence. And it created this method with an insight that had eluded others. Specifically, ITRI based its innovative method on the insight that *mismatches* between the forward and reverse strands of the chemically altered DNA could be used to identify modified bases. The prior art knew that mismatches sometimes occurred between forward and reverse strand sequences, but did not use them to identify modified bases. That is, researchers occasionally observed such mismatches when the wrong base was detected in one strand's sequence and the correct base was detected in the other. Because a mismatch is a violation of the

Watson-Crick rules, the mismatching data would be considered unreliable and thus thrown out or overruled. Thus, until ITRI's invention, using mismatches to identify modified bases was unknown. Indeed, the prior art relied on "consensus sequencing," which required redundancy to generate a sufficient number of matches in the sequencing data before results were considered reliable.

The Board below erred in failing to appreciate ITRI's insight regarding mismatches. Instead, the Board took for granted that because the prior art had known that mismatches can occur, the prior art also must have known they can be used to locate modified bases. But the only source for the latter point—that mismatches were actually useful and not indicators of unreliability to be discarded—came solely from ITRI's application.

This is true for the Board's two principal decisions, where it declared PacBio to be the senior party and held that all of ITRI's claims were obvious. But PacBio's priority application and the prior art relied on consensus sequencing (which looks only at matching data), and neither indicated in any way that mismatches were useful for any reason, let alone locating modified bases.

The Board's rejections of ITRI's other claims were also improper because elements that the Board found obvious were not even mentioned in the references the Board applied. In one case, the claims recited ITRI's novel way to confirm sequencing-data reliability regardless of the presence of mismatches by adding

inserts to the DNA with already-known and thus easily cross-checked sequences. In another case, the claims recite a discriminating analog that paired preferentially with one or the other of a base or its modified form. Finally, after improperly rejecting all of ITRI's claims, the Board apparently shielded PacBio from the same prior art by crediting an "information-and-belief" declaration asserting common ownership that was contradicted by the inventorship of record.

In sum, the Board committed multiple errors below, most of them traceable to that most common of legal errors in patent law, impermissible hindsight. Accordingly, the Board's decisions should be reversed and the interference remanded for further proceedings conducted in accordance with the legal standards this Court's precedents require.

IV. STATEMENT OF THE CASE

A. Nature of the Case

DNA carries genetic information in the sequence of four nucleotide bases, adenine, guanine, cytosine, and thymine, conventionally abbreviated as A, G, C, and T, respectively. DNA also carries "epigenetic" information when the bases are chemically modified. The modified bases have important effects on regulating gene activity. Locating the modified bases and understanding the epigenetic information they convey is of great interest.

Bases in a DNA double helix generally match according to the Watson-Crick rules—A matches with T and C with G. ITRI's inventors located certain modified bases in single DNA molecules by using mismatches (Watson-Crick violations) between opposite strands of the same double stranded DNA sample. This approach differed from prior methods, which used a second reference sequence to identify modified bases instead.

The sole Count in this interference is a method of sequencing a single DNA molecule sample and determining the position of modified bases in that sample based on mismatches. The method takes advantage of the fact that DNA in the genome occurs in double-stranded form, with the two strands twisted around each other in a double helix, with the bases in the strands matched according to the Watson-Crick rules. According to the Count, a circular pair-locked molecule (CPLM) is formed from a double-stranded DNA sample by locking the two strands together end-to-end. Thus, the CPLM is a single DNA molecule that contains what were previously separate, but complementary, strands. In the invention of the Count, there is at least one modified base in the CPLM that is mismatched (paired with a base that it does not normally pair with). The mismatch between the sequences of the forward and reverse strands is used to identify the modified base and its position.

B. Course of Proceedings and Disposition Below

The Count is claim 24 of ITRI's U.S. Patent No. 8,486,630. PacBio copied ITRI's claims into one of its involved applications (13/633,673) to provoke an interference with the '630 patent. Unable to convince the examiner that it had written-description support for the copied claims in its earliest priority application, PacBio narrowed one claim and canceled the rest. The Board declared Interference No. 105,970 between ITRI's '630 patent and PacBio's Application No. 13/633,673. A111-14. A second PacBio application, No. 13/930,178, later entered the interference. A250-52. All claims of both parties corresponded to the Count. A111-15; A250-52.

The Board accorded PacBio the benefit of several applications, the earliest of which was Application No. 61/201,551, filed December 11, 2008. A115. The Board accorded ITRI the benefit of Application No. 61/167,313, filed April 7, 2009. *Id*. Based on these filing dates, the Board designated PacBio the senior party. A112-14.

The parties filed motions including ITRI's motion to rescind the benefit of PacBio's '551 application (the "benefit motion") and PacBio's motion alleging that all the claims of the '630 patent were obvious (the "obviousness motion"). PacBio's only benefit date earlier than ITRI's benefit date of April 7, 2009, came

from the ’551 application. A115. Thus, granting the benefit motion would have made PacBio the junior party and allowed ITRI to remain in the interference.

The Board granted PacBio’s obviousness motion and denied ITRI’s benefit motion. On appeal, ITRI challenges the Board’s determination that PacBio had a written description of an embodiment of the Count in its ’551 application and that ITRI’s claims would have been obvious. ITRI also challenges the Board’s failure to hold PacBio’s claims unpatentable as required by 37 C.F.R. § 41.207(c) given PacBio’s failure to rebut the presumption that the asserted prior art rendered its own claims unpatentable.

V. STATEMENT OF FACTS

A. Technological Background

1. Watson-Crick base pairing

DNA in the genome occurs in double-stranded form, with the two strands wound around each other in a double helix. A81[Fig.14]; A88[8:51-55]. The two strands are referred to as “forward” and “reverse” strands. *See, e.g.*, A81[Fig.14]; A88[8:51-55]. The forward and reverse strands usually contain information that is complementary and redundant—one can deduce the sequence of one strand from the sequence of the other strand using Watson-Crick base pairing rules (A pairs with T, and C pairs with G). *See, e.g.*, A81[Fig.14]; A88[8:48-58]. The pairing rules result from the paired bases having complementary shapes. Within a DNA

double helix, A fits best with T, and C fits best with G, much like pieces of a puzzle. Thus, an A in one strand indicates a corresponding T on the other and vice versa, and a G in one strand indicates a C on the other and vice versa. A96[23:12-15]. During DNA synthesis, a strand is separated from its complement and an enzyme constructs a new strand by piecing nucleotides together based on their shapes, one at a time, using the separated strand as a template. For example, an A is added to the new strand opposite a T on the template.

2. Modified bases

DNA can also carry information "epigenetically," using chemically modified bases. A85[2:14-18]. There is great interest in determining the existence and positions of modified bases because of their biological roles. *See*, *e.g.*, A85[2:14-18]. Uracil (U) and 5-methylcytosine (^{m}C) are two examples of modified bases that can occur in DNA. A95-6[22:64-23:2]. When present, U pairs with A and thus is similar to T. A835[¶21]. Pairing of C and ^{m}C, in contrast, is the same–both pair with G. A72[Fig.5A].

3. Sequencing-by-synthesis

When the parties' applications were filed, genetic sequence data from single DNA molecules could be determined using a "sequencing-by-synthesis" approach. A917-18[¶16]. Sequencing-by-synthesis involves monitoring the order of nucleotide addition while synthesizing a complementary nucleic acid strand. A92-

93[16:4-17:29]. Researchers can determine the identity of the nucleotides in the original molecule using the Watson-Crick base-pairing rules. A835-36[¶23]. For example, detecting addition of an A to a growing DNA strand reveals a T at the complementary position in the target template strand. A81[Fig.14]; A88[8:48-58]; A835-36[¶23].

But there are limits to sequencing-by-synthesis when modified bases are involved. In particular, applying the Watson-Crick rules to sequencing-by-synthesis results cannot identify whether or where chemical modifications occur in the sequence unless more is done. For example, in sequencing-by-synthesis, the addition of an A to a growing DNA strand does not indicate whether the complementary base on the template strand being sequenced was a T (normal base) or U (modified), nor does the addition of a G indicate whether the complementary base on the template strand was a C (normal) or ^{m}C (modified). In other words, one needs more information than which base was added to distinguish Ts from Us or Cs from ^{m}Cs.

A prior-art technique for distinguishing C from ^{m}C, and thus determining the ^{m}C positions begins with a reference sequence. DNA with the same sequence is then subjected to bisulfite treatment, which converts C into U but does not affect ^{m}C or other bases. A96[23:57-59]; A89[9:1-2]. The prior-art technique then sequences the bisulfite-treated DNA. A846-47[¶60]. Comparing the treated

sequence to the reference sequence reveals where conversion occurred because the locations that originally had a C now have a U, and these bases behave differently during sequencing. In contrast, the ^{m}C positions show no change.

All these prior-art sequencing methods are based on the assumption that sequencing data will be accurate. But sequencing is typically not 100% accurate. So researchers would occasionally test the reliability of their sequence data. A common way to do this was, after sequencing, comparing the forward and reverse strands. If the bases on the forward and reverse strands match, researchers can be confident that the sequencing method is reliable. *See, e.g.*, A847[¶60]; A860[1:21-27] ("If the bases on both strands complement each other as expected, then this helps to confirm the reliability of the sequence information."). Mismatches between the forward and reverse sequences indicated unreliability, so researchers would ignore or discard results with such mismatches or overrule the mismatches with additional data. A1015[¶49].

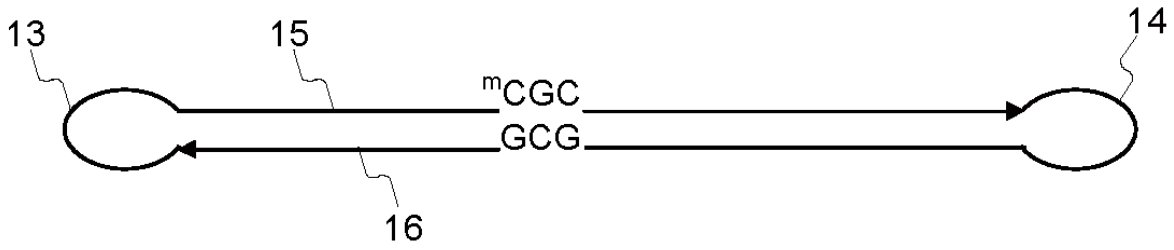
B. ITRI's 2009 Inventions

ITRI's inventors developed better ways to identify modified bases in single DNA molecules. The first involves comparing forward and reverse DNA strands to look for mismatches. Unlike the prior art, the inventors used the mismatches to locate modified bases. The second involves the use of "discriminating analogs,"

which are nucleotide analogs that are capable of distinguishing between normal and modified bases.

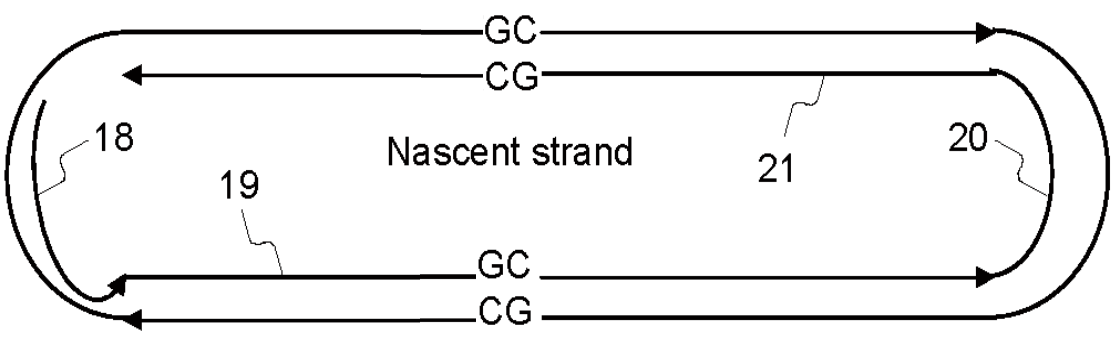
1. Using mismatches to locate modified bases

Forward and reverse DNA strands are separable but can be locked together at both ends to produce a circular pair-locked molecule (CPLM)—a single circular molecule that contains within it both the forward and reverse strands. A101[34:23-32]. The figure below shows an exemplary CPLM structure:



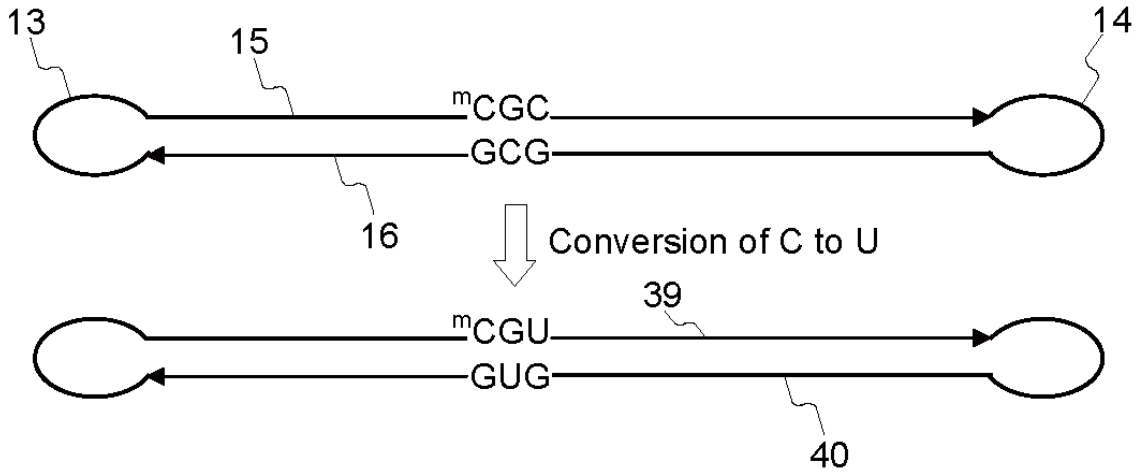
A72[Fig.5A]. In the figure, the forward and reverse strands of a DNA sample 15 and 16 are locked together with DNA inserts 13 and 14.

Sequencing-by-synthesis of a CPLM can then generate sequence data for both of the original forward and reverse complementary strands because the forward and reverse strands are part of the same molecule, A101[34:23-32], as shown below:



Cf. A73[Fig.6A].[1] Here, sequencing-by-synthesis is in progress starting from 18, producing a DNA molecule that contains a sequence complementary to the reverse strand (19), a linker (20), and the forward strand (21).

In such sequencing, certain modified bases may be "mismatched" in the original forward and reverse strands, i.e., paired with a different base than the one dictated by the Watson-Crick pairing rules. For example, when subjected to bisulfite conversion, cytosine (C) becomes uracil (U), creating U-G pairings where there used to be C-G pairings. A72[Fig.5B]; A96[23:2-6]. But, because U normally pairs to A and not G, these U-G pairings are "mismatched."



A72[Fig.5B].

[1] Certain details in Fig. 6A were omitted for clarity.

But the bisulfite conversion reaction does not affect 5-methylcytosine (^{m}C). A89[9:1-2]. Thus, bisulfite conversion of a molecule containing both C and ^{m}C alters only C. *See, e.g.*, A72[Fig.5B] (illustrating conversion of C but not ^{m}C). Therefore, ^{m}C generally *does not* occur in mismatches, but U *does* occur in mismatches when bisulfite conversion is used. For this reason, the term "mismatched modified base" will generally include U but not ^{m}C.

ITRI's '630 patent describes sequencing a CPLM having a mismatched modified base by producing sequence data that shows a *mismatch* between the forward and reverse strand sequences in at least one position, and using the mismatch to determine the position of the modified base. A97[25:52-26:23]. For example, with U-G pairing, the sequencing data would show a mismatch because an *A* would be added opposite the U and a *C* would be added opposite the G, revealing the mismatched U-G pairing in the forward and reverse strands of the CPLM. A97[25:60-63]; A838[¶29].

This method differs from the prior art because the prior art did not use a mismatch between forward and reverse strands to locate a modified base. Instead, it compared the sequence of a treated strand to the reference sequence of the untreated version of the same strand, thus requiring possession (usually by separate sequencing) of the untreated reference sequence. In ITRI's methods, which are reflected in claims 1-26, one compares the sequences of the forward and reverse

strands, and determines that a modified base is present at positions where there is a mismatch. A838[¶30]; A107-08[45:24-48:14].

2. Confirming the reliability of data including mismatches with known "insert" sequences

Because the inventors developed a way to use mismatches to find modified bases and not necessarily as indicators of unreliability, they also worked on new ways to confirm sequence reliability. Specifically, they determined they could take advantage of the known sequences of the DNA inserts they used to lock the forward and reverse strands together. By studying the insert sequence results they could decide whether to accept or reject sequencing results for the forward and reverse strands. In essence, the accuracy of the known insert sequence serves as a proxy for the accuracy of the to-be-determined sequence of the forward and reverse strands.

This approach represents a way of improving sequence reliability different from using redundant sequencing at individual positions of the forward and reverse strands. Differences from a known insert sequence necessarily show an actual error, thus avoiding the need to predict errors from redundant sequencing information of strands with unknown sequences. The insert sequences are known, and the accuracy of the sequencing data for the inserts is determined by calculating scores for at least four sequences of the inserts by comparing the obtained and

known sequences. A108[47:15-18]. The scores reflect the degree of closeness of the obtained insert sequences to the known insert sequence. A93[18:52-57].

3. Using discriminating analogs to locate modified bases

The ’630 patent describes another way to locate modified bases that does not rely on using a mismatch. Instead, this alternative embodiment employs a “discriminating nucleotide analog” to determine modified base positions. Specifically, the ’630 patent states that a “discriminating analog” is one that “pairs preferentially with one but not the other of the base and its modified form.” A96[24:15-19]. The discriminating nucleotide analogs in ITRI’s patent are species of “base-linked” analogs, meaning that the analog contains chemical group(s) connected to its nucleotide base that result in the discrimination. A96-97[24:42-25:24]; A1022-23[¶¶68-69].

C. PacBio’s 2008 Research Activity

Around the time ITRI was developing its methods for locating modified bases, PacBio was actively developing methods for improving sequencing accuracy (and thus reliability) through the use of redundancy or “consensus sequencing.” A1014-15[¶¶47-48].

1. PacBio’s work on “consensus sequencing”

U.S. Patent No. 8,153,375, which was based on provisional applications filed in March and September 2008, reflects PacBio’s activities at that time. A684. So does a presentation PacBio made at an October 2008 Cold Springs Harbor

Personal Genomes Meeting ("*Personal Genomes*"). A937. As indicated in the '375 patent and *Personal Genomes*, PacBio was studying CPLMs and single-molecule sequencing. A718[31:35-32:67]; A938; A943; A966-70. But, at least in early 2008, PacBio was not investigating methods for locating modified bases, let alone determining the modified-base position by comparing sequences of forward and reverse strands. A1014[¶¶46-47]; *see also* A260 (PacBio's admission that neither the '375 patent nor *Personal Genomes* "explicitly discloses the use of the template to determine the position of a modified base, or altering the base pairing specificity of a specific type of base in the template and thereafter determining the position of the modified base").

Instead, both describe methods for improving sequencing reliability by comparing sequence data of forward and reverse strands to determine where they show consensus, i.e., where they agree. A707-08[9:64-10:2] ('375 patent: describing "consensus sequence determination through the sequencing of both the sense and antisense strand"); A980-81 (*Personal Genomes*: describing "consensus accuracy"); A1014[¶47]. In particular, PacBio improved sequencing reliability through redundancy, i.e., sequencing the same forward and reverse strands multiple times. A707[9:49-58] ('375 patent: "sequencing a given template provides duplicative or replicate data of the sequence information obtained, and thereby improves accuracy over linear templates by providing multiple reads for a given

template . . . that can be used to derive consensus sequence data"); A980-81 (*Personal Genomes*: "Consensus Accuracy Increases with Coverage," i.e., redundancy); A1014-15[¶¶47-49].

2. PacBio's work on labeled nucleotides

PacBio was also working on a way to improve sequencing-by-synthesis by avoiding the use of "base-labeled" or "base-linked" nucleotides, that is, nucleotide analogs in which a fluorescent label is linked to one of the nucleotide bases such as C or G, which PacBio reported to be "problematic." A946-47; A1022-23[¶69]. Specifically, base-linked nucleotide analogs caused various problems including impeding the reaction by inhibiting the polymerase enzyme used to add bases to the extending strand. A946 (recognizing that "Base-Labeled Nucleotides are Problematic" because they "[i]nhibit[] enzyme" and cause other problems specific to the fluorescent label). As a solution to the problems associated with base-linked nucleotide analogs, *Personal Genomes* taught instead using labeled nucleotides having a fluorophore that was not linked to the base of the nucleotide. A946-47; A1022-23[¶69]; A442[¶127]. These labeled nucleotides did not, however, discriminate between a base and its modified form.

3. PacBio's work on modified bases

In December 2008, PacBio filed the '551 application. It is directed to identification of modified bases in single DNA molecules, but not by looking for

mismatches in the molecules' forward and reverse strands. To the contrary, the '551 application expressly disclaims the use of any method that "rel[ies] on the similarity of uracil to thymine," which is exactly how ITRI's method works for samples subjected to bisulfite treatment—mismatches are determined based on the similarity of uracil to thymine. A920[¶23]. Instead, the '551 application states that "the methods herein directly detect the modified base" (*Id.*) by analyzing a statistical set (more than a single round) of sequencing data (A922-25[¶¶30, 33, 36]; A839[¶33]). More specifically, the '551 application describes approaches for determining the position of a modified base using a distinct signal, such as a different misincorporation (error) frequency or an effect on incorporation data such as altered kinetics of incorporation (e.g., stalled or slowed base incorporation), which is observed during the sequencing-by-synthesis of a strand containing a modified base. A917-19[¶¶16, 17]; A921-22[¶¶28, 29]; A839[¶33]. Thus, determining a modified base position according to the '551 application must rely on a distinct signal arising from the sequencing of the modified base that does not occur with an unmodified base. A839[¶33].

Incorporating by reference PacBio's earlier applications directed to improving the accuracy of sequencing data, the '551 application states that the consensus-sequencing method of the '375 patent can be used to "further validate" the results obtained using the error-frequency or stalling methods it describes:

> In certain embodiments, redundant sequence information can be generated by sequencing the same template nucleic acid multiple times by, e.g., sequencing the same template nucleic acid molecule repeatedly or by sequencing multiple copies of the same template nucleic acid. Methods for generating redundant sequence information are provided, e.g., in U.S. Pub. Patent Application 20060063264, filed September 16, 2005; U.S.S.N. 61/094,837, filed September 5, 2008; and U.S.S.N. 61/099,696, filed September 24, 2008, all of which are incorporated herein by reference in their entireties for all purposes. Such redundant sequence information can provide additional information for discriminating modified from unmodified nucleotides. For example, sequence reads from the sense or "forward" strand can be compared to sequence reads from the antisense or "reverse" strand for the same nucleic acid template to further validate the existence of one or more modified bases in the template nucleic acid.

A918[¶17] (incorporating by reference, *inter alia*, "U.S.S.N. 61/099,696," an application on which the '375 patent was based).

The passage quoted above discusses the possibility of redundant sequencing, possibly including using forward and reverse strands, according to the various incorporated applications. The application then moves on to another, different method that uses redundant sequencing to calculate misincorporation rate at a position in one strand. *See id*.

The method that uses misincorporation rate does *not* involve a mismatch in the double-stranded sample. The misincorporation rate measures how often a wrong base is added during sequencing, which occurs regardless of whether the

modified base is mismatched to its reverse strand in the sample or whether the reverse strand is even sequenced at all. A839-40[¶¶33-34]; A850-51[¶73]; A918-19[¶17] (discussing detection of "5-methylcytosine"—which is not mismatched—using misincorporation rate in "the newly synthesized strand"). For example, during sequencing, an A should be added opposite a U, but sometimes a G is added opposite a U. *Id.* Thus, the '551 application taught that a U can be detected by determining the rate at which G's are added ("misincorporated") at a position during the sequencing reaction and that ^{m}C can also be detected in this way. *Id*.

Misincorporation, which occurs when the wrong base is added *during* sequencing-by-synthesis, is different from a mismatched modified base in the CPLM, which exists *before* sequencing begins and is detected without misincorporation. *E.g.*, A72[Fig.5B], A73[Fig.6A]. Thus, PacBio's '551 application's disclosure of misincorporation rate uses data from a single strand instead of the Count's method, which requires looking for mismatches by comparing forward and reverse strands.

D. The Interference

1. The Count requires the use of mismatches to determine modified base positions

The Count of the interference, which was claim 24 of ITRI's '630 patent, requires sequencing a double-stranded nucleic acid sample and determining the position of at least one modified base in the sequence based on detecting a

mismatch between the forward and reverse sequences of the CPLM. A110-18; A838[¶30].

In step (a) of the Count, a CPLM is formed from a nucleic acid sample. A114; A837[¶ 27].

In step (b), sequence data of the CPLM is obtained via single-molecule sequencing. Within that sequence of the CPLM are the sequences of both the forward and reverse strands of the sample. A114; A837[¶28].

In step (c), the sequence of the sample, including the position of the at least one modified base in the sample, is determined by comparing the sequences of the forward and reverse strands of the CPLM. A114; A838[¶¶ 29-30]. Step (c) further recites that "at least one modified base in the double-stranded nucleic sample is paired with a base having a base pairing specificity different from its preferred partner base," i.e., does not conform to the Watson-Crick base-pairing rules, and is thus mismatched. A114; A837[¶26]; A838[¶29]; A96[23:12-33].

The Board accepted ITRI's claim construction based on intrinsic evidence and expert testimony, under which step (c) of the Count requires using a mismatch between the forward and reverse strand sequences of a CPLM to determine the position of the modified base. *See, e.g.*, A40[13-18] (discussing "looking for guanine-thymine mismatches as evidence of non-methylated cytosine in a forward or reverse strand locus"); A97[25:60-63] ('630 patent: "A position at which there

is disagreement among the repeats (e.g., the position labeled 41 in FIG. 6B) signifies that a base in the nucleic acid sample at that position underwent alteration of its base pairing specificity."); A838[¶30] (expert testimony).

Importantly, when the modified base being detected is uracil, the method of the Count depends on the similarity of uracil to thymine. A851[¶75]. The detection of uracil requires that it pair with adenine during sequencing, which is similar to how thymine would pair with adenine. *Id.* Indeed, PacBio admitted that pairing to adenine is a *similarity* shared by uracil and thymine. A319[¶60].

2. The Board decided that PacBio's '551 application discloses an embodiment of the Count

In challenging PacBio's claim for priority to the '551 application, ITRI argued that it provided no written-description support for the Count. A1029. Specifically, ITRI argued two grounds: first, nothing in the '551 application disclosed comparing forward and reverse strand sequences that show a mismatch at a modified base position and then using that mismatch to determine the modified base position (step (c) of the Count), and second, the '551 application expressly disclaimed methods, like ITRI's, that "rely[] on the similarity of uracil to thymine." A1042-43.

a. ITRI's first ground: there is no disclosure in the '551 application where mismatches in the forward and reverse strands are used to locate modified bases

There is only a single mention of using the forward and reverse strands of the sequenced DNA in the '551 application: one sentence in paragraph 17. But paragraph 17 incorporates by reference an application the '375 patent was based on and merely describes how the '375 patent's redundant/consensus sequencing approach can be used to "further validate," i.e., confirm the accuracy of, the results obtained by the signal-based methods described in the '551 application. *See supra* pages 16-20. A consensus sequence—as its name implies—looks for *agreements* (matches) in the sequencing data as an indication of reliability and discards disagreements (mismatches) as errors. *Id.*

To support these arguments, ITRI submitted the declaration of Dr. Shawn Levy, whom the Board recognized as an expert qualified to provide testimony on these matters. A16. Dr. Levy explained that the "redundant sequence information" described in paragraph 17 was a reference "to the known technique of validating sequence of one strand based on the matching (i.e., redundant) sequence of the reverse strand." A848[¶65-66]. Thus, he explained, paragraph 17 taught that the accuracy of results can be confirmed with "consensus sequencing" (*Id.*) of the type that PacBio was touting to the scientific community at large during this timeframe (A980-81 ("Consensus Accuracy Increases With Coverage")).

b. ITRI's second ground: the '551 application disclaims reliance on uracil's similarity to thymine, which is a necessary aspect of the Count

The '551 application states: "Uracil can be found in DNA . . . as the result of bisulfite-conversion of cytosine in a common protocol used to discriminate methylated cytosine through DNA sequencing. *Unlike this common protocol*, the methods herein directly detect the modified base *rather than relying on the similarity of uracil to thymine*." A920[¶23] (emphasis added). Instead, rather than using this similarity with thymine (pairing with adenine) to identify the uracil, the '551 application discloses a distinctly different approach that relies on *differences* between uracil and thymine, such as a higher error rate or slowed incorporation, to *distinguish* uracil from thymine. *See supra* pages 18-20.

PacBio admitted that one *similarity* shared by uracil and thymine is that they both normally pair to adenine. A319[1-4]. Therefore, ITRI argued that the '551 application excludes any method that relies on uracil's and thymine's similar propensities to pair with adenine, including the method of the Count. Specifically, according to the Count, a modified base must be mismatched in the CPLM (e.g., uracil when bisulfite conversion was used). When the modified base is uracil, the Count requires the use of the similarity between uracil and thymine, i.e., that they both pair to adenine during the sequencing reaction, to detect the mismatch. Accordingly, ITRI concluded, when uracil is the modified base being detected, the

invention relies on a similarly between uracil and thymine, and thus has been disclaimed by the ’551 application. A1042-43; A851[¶74]. And the ’551 application did not disclose any mismatched modified base except uracil.[2] A850[¶78]. Thus, excluding reliance on the similarity of uracil to thymine precluded description of the only possible embodiment of the Count that the ’551 application otherwise could have described.

c. The Board’s analysis

In considering the teachings of the ’551 application, the Board acknowledged and did not disagree with Dr. Levy’s factual testimony that a consensus sequence relies on *agreement* in order to confirm the location of a modified base. A53[26-31]. Nor did the Board disagree that the ’551 application disclaimed any method relying on the similarity of uracil and thymine. Indeed, the Board never even mentioned that issue. Also, as ITRI asserted, the ’551 application failed to expressly disclose an embodiment of the Count, and neither PacBio nor the Board cited any disclosure supporting an express disclosure. Thus, both PacBio and the Board had to rely upon an inherent disclosure to find written description of an embodiment of the Count. To support an inherent disclosure, the specification

[2] The Examiner agreed that other mismatched modified bases were not described, as shown by the requirement to limit Pacific Biosciences’ claim by reciting bisulfite treatment. A1056.

must necessarily disclose an embodiment of the Count, not just show the possibility.

The Board, however, did not find that one skilled in the art would consider the '551 application as *necessarily disclosing* using a mismatch to find a modified base. Instead, the Board framed the question as being whether "the '551 application necessarily requires that the analysis of base pair [*sic*] *only validate that the base pairs agree with the Crick-Watson rules*." A54 (emphasis added). In other words, instead of focusing on whether the '551 application necessarily *disclosed* step (c) of the Count (using the mismatches to identify modified base locations), the Board focused only on whether the '551 application necessarily *excluded* it.

The Board found support for the Count in the '551 application (and paragraph 17 in particular) because "a person of ordinary skill would expect to see mismatched uracil-guanine base pairs in a DNA sample strand that had been modified by bisulfite conversion." A54[12-14]. In reaching this conclusion, the Board did not cite any supporting expert testimony. *See* A53-55. Instead, the Board relied solely on paragraph 17 for support (*id*.), even though paragraph 17 says nothing about mismatches, let alone using them to identify modified base locations.

The Board filled this gap in paragraph 17 with what it deemed to be "*a priori*" knowledge concerning uracil's propensity to pair with adenine, a property uracil shares with thymine, despite the '551 application's disclaimer of any method "relying on the similarity of uracil to thymine"—"that bisulfite converted uracil would be expected to be paired, incorrectly, with the guanine that would normally be paired with the pre-converted cytosine." A54[7-12].

Further, the Board never discussed whether the '551 application *necessarily disclosed* using mismatches to locate modified bases. Instead, it focused on what would be "reasonable," what one of ordinary skill would "recognize," and what one of ordinary skill would "understand":

> [G]iven the relatively high level of skill of the average artisan, as we have defined it, it is not *reasonable* to believe that such a person would not *recognize* a mismatch occurring between a modified base, i.e., cytosine converted to uracil by bisulfite treatment (as taught by the '551 application) paired with guanine, and not *understand* from that mismatch that "at least one modified base in the double-stranded nucleic sample is paired with a base having a base pairing specificity different from its preferred partner base" as required by the disputed limitation.

A54[1-7] (emphasis added).

3. The claims of ITRI's '630 patent, the prior art, and the Board's obviousness analyses

a. Facts relating to claims 1-26

The Board held that claims 1-26 would have been obvious over U.S. Patent No. 8,153,375 in view of Laird *et al.*[3] ("*Laird*") and Matsumura *et al.*[4] ("*Matsumura*"). A26-27, A43.

Claims 1 and 26 of ITRI's '630 patent comprise elements that, for purposes of this appeal, do not differ materially from claim 24, which is identical to the Count (discussed above). A1011-13[¶¶38, 39, 41]. Both claims 1 and 26 recite an additional limitation not included in claim 24 of "altering the base-pairing specificity of bases of a specific type in the [CPLM]," which produces altered bases that are mismatched in the CPLM. A1011[¶¶39-40]; A107[45:36-38]; A108[48:4-5]. The modified base position is determined by steps similar to those in claim 24. A107[45:44-47]; A108[48:10-14]. In other words, claim 24 recites an already-present mismatched modified base, whereas claims 1 and 26 recite steps (such as bisulfite treatment) that create mismatched bases prior to detecting them.

---

[3] Laird et al., (2004) Hairpin-bisulfite PCR: Assessing epigenetic methylation patterns on complementary strands of individual DNA molecules, *Proc. Natl. Acad. Sci. USA* 101(1): 204-09.

[4] Matsumura et al., (2007) Photochemical Transition of 5-methylcytosine to Thymine by DNA Photoligation, *Nucleic Acids Symp. Ser*. 51: 233-34.

Thus, as the Board found, one of ordinary skill would interpret claims 1 and 26 as requiring that a mismatch between the sequence data of the forward and reverse strands of a CPLM is used to determine the position of a modified base, much as with claim 24. A1011-13[¶¶39-44].

i. U.S. Patent No. 8,153,375

The ’375 patent discloses CPLMs and single molecule sequencing, but does not disclose determining a modified base position by comparing sequences of forward and reverse strands. *See supra* pages 16-18. Instead, the ’375 patent discloses the use of redundancy, or consensus sequences, to verify the reliability of sequencing data. *Id.*

ii. *Laird*

Laird is a 2004 publication focusing on “the fidelity of methylation transmission” from a cell to its progeny for a known gene, *FMR1*. A722[Abstract]. “Methylation transmission” refers to inheritance, i.e., whether, after a parent cell copies its DNA and divides, both new resulting cells receive the same methylation pattern (i.e., whether the 5-methylcytosine appears in the same locations in the newly created strands).

Before determining whether the new cells receive the same methylation pattern, *Laird* located the 5-methylcytosine sites in the conventional way, by comparing an untreated reference strand sequence to the sequence of a bisulfite-

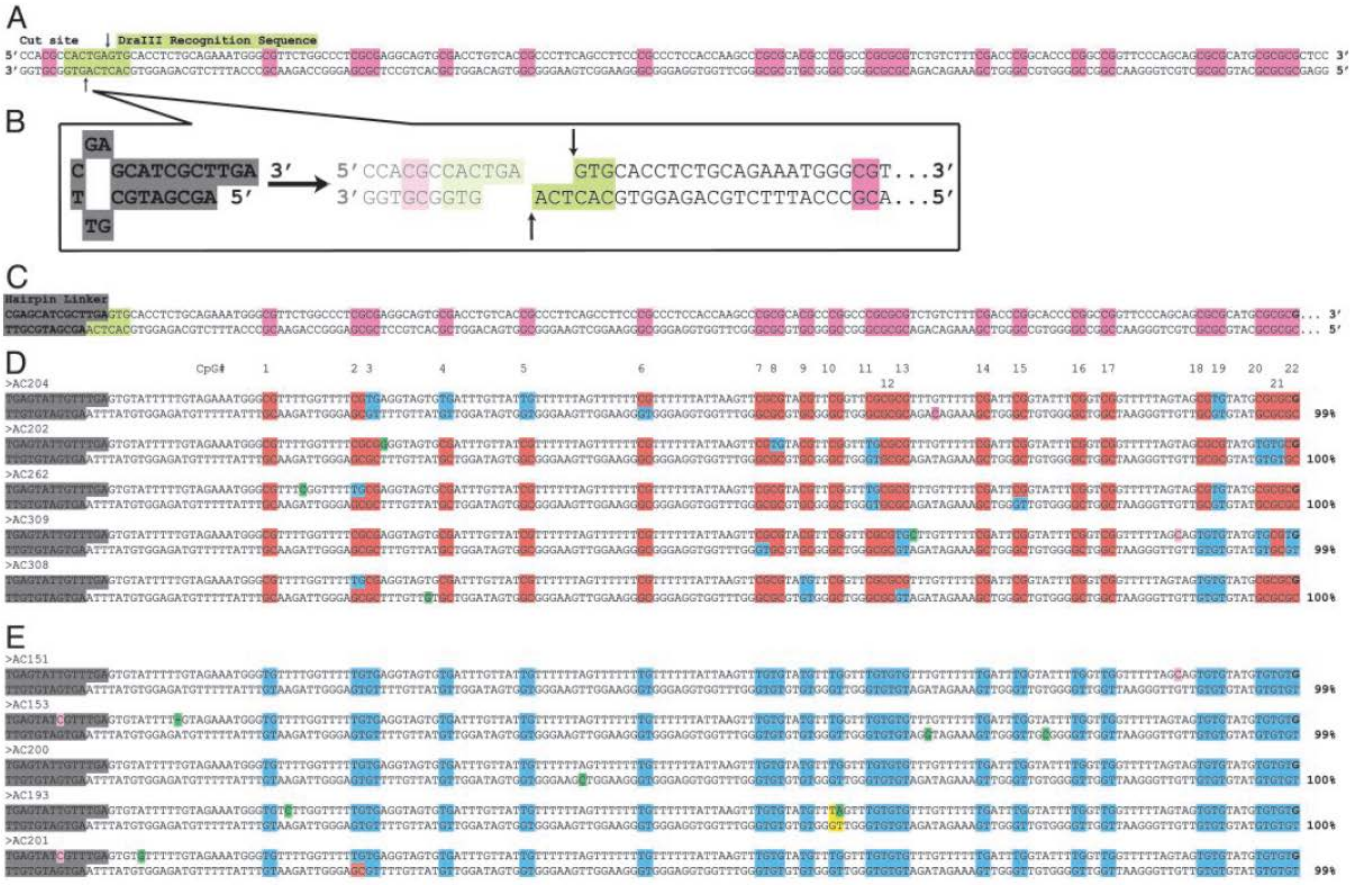
treated version of the same strand. *See supra* pages 9-11. Specifically, *Laird* reported the "frequency of methylation" in one strand after bisulfite treatment. A724. Obtaining this frequency requires dividing (i) the number of cytosines in the strand before treatment by (ii) the number of cytosines in that strand after treatment. A1018-19[¶59]. Thus, Laird compared the strand before and after treatment, which was the conventional way to determine modified base positions after bisulfite treatment. A1018-19[¶¶59-60]; A846-47[¶60]. And *Laird* never mentions looking for mismatches to identify a modified base position.

Only *after* identifying the 5-methylcytosine positions separately for each strand did *Laird*, for the first time, compare the forward and reverse strands to determine the extent of *symmetry* to assess whether methylation was inherited. Symmetry refers to whether two cytosines on opposite strands are both methylated or both unmethylated, as illustrated below.

- -^{m}C - G- -	- -^{m}C - G- -	- - C - G- -
- - G -^{m}C- -	- - G - C- -	- - G - C- -
Symmetric (methylated)	Asymmetric	Symmetric (unmethylated)

Laird analyzed symmetry by observing where the already-identified modified base positions in one strand lined up with the already-identified modified base positions in the other strand. A724-25. That is, the opposite strand did not come into play until after the step of determining the position of a modified base.

Figures 2D and 2E of *Laird* show this analysis of symmetry of the already-identified methylated (red) and unmethylated (blue) positions:



A725.

Although you can see mismatches if you look very closely at the blue-highlighted portions of the treated strands (with a magnifying glass or by zooming in on an electronic copy) in Figure 2 of *Laird*, *Laird* never mentions mismatches, let alone describes or suggests looking for them to identify modified base positions. A846-47[¶60]. Nor is there any indication that *Laird* used any information from the opposite strand in making the determination of whether any

given cytosine was methylated (and thus bisulfite-resistant) or unmethylated (and thus bisulfite-sensitive, meaning that it underwent bisulfite conversion to uracil).

iii. *Matsumura*

Like *Laird*, *Matsumura* disclosed a method of converting bases to determine cytosine methylation. A728[Abstract]; A1018-20[¶¶59,61]. Specifically, *Matsumura* converted ^{m}C to T by ultraviolet irradiation. In *Matsamura*'s study, the substrate was a single-stranded molecule that was only six nucleotides long. A728-29; A1019-20[¶61]. *Matsumura* compared data from the single-stranded substrate before and after ultraviolet radiation to determine that conversion had occurred. A729[Fig.2]; A1019-20[¶61]. PacBio admitted that *Matsumura* disclosed determining that a conversion occurred using data from a small single-stranded molecule before and after the conversion. A440[¶114].

In other words, *Matsumura*'s substrate was single-stranded, so there was no reverse strand. And *Matsumura*'s comparison of before-and-after data for one strand was similar to the prior art's (and *Laird*'s) comparison of the same strand before and after treatment. A1019-20[¶61].

iv. The Board's analysis

The Board held that it would have been obvious to modify the '375 patent in view of *Laird*[5] to reach claims 1-26. A42-43. The Board based this conclusion on its finding that "*Laird* explicitly teaches looking for guanine-thymine mismatches as evidence of a non-methylated cytosine in a forward or reverse strand locus." A40[35:16-18]. Even though the Board found that *Laird* "*explicitly*" teaches using mismatches to locate modified base positions, there is no mention of mismatches in *Laird*, including in the sections quoted by the Board. Instead, the sections quoted by the Board merely described the well-known effect of bisulfite treatment on cytosine: it converts cytosine to uracil, but does not affect 5-methylcytosine. A38-43.

b. Facts relating to claim 23

The Board stated that ITRI did not argue any claims separately. A31[n.29]. However, ITRI presented separate arguments, including for claim 23. A342-43.

Claim 23 depends from claim 1 and recites additional limitations including using known insert sequences to lock the forward and reverse strands together, and using the accuracy of the sequencing results for those insert sequences to decide whether to accept or reject sequencing results for the forward and reverse strands.

[5] The Board's obviousness analysis does not cite any specific teachings in *Matsumura*. *See* A38-43.

See supra pages 15-16. In the claim 23 method, the sequencing data comprises at least two copies of the sequence of the CPLM, each copy comprising sequences of first and second "insert-sample units," i.e., the sequences of both strands of the sample ("oppositely oriented repeats of the sequence of the nucleic acid sample") and the adjacent inserts. A108[47:6-9].

Claim 23 then recites "accepting or rejecting at least four of the repeats of the sequence of the nucleic acid sample contained in the sequence data according to the scores of one or both of the sequences of the inserts" adjacent to the sample sequences. A108[47:19-23]. The scores used to accept or reject repeats of the *sample* sequence are calculated by comparing the obtained *insert* sequences to their known sequences, A108[47:6-18]. Thus, sample sequences are accepted or rejected based on results from the insert sequence, which *differ* from the sample sequences.

The Board's decision that claim 23 would have been obvious stated that the '375 patent "explicitly teaches determining a consensus sequence *at a position* from 'multiple reads of *that position* of sequence.'" A41 (quoting '375 patent) (emphasis added). But the Board did not discuss the claim requirement of using a score of one sequence to accept or reject a different sequence. The Board further stated that "*Laird* teaches determining sequences and agreement/disagreement of base pairs at given loci." A41. But the Board never found that *Laird* used scores of any type to accept or reject sample sequences.

To perform the method of claim 23, one must obtain and use a score for insert sequences adjacent to the sample sequences to accept or reject sample sequences. *See supra* pages 15-16. The Board did not point to evidence that one of ordinary skill would have accepted or rejected a sample sequence *according to the score of a different insert sequence*. To the extent any scoring occurred in the prior art, it was of the same sequence, not of an insert. A41; A708[11:40-49]. Neither PacBio's motion nor the references it cited mentioned accepting or rejecting repeats of a *sample* according to the score of an *insert with a known sequence* in its argument about claim 23. A708[11:40-53]; A610-11[¶¶35]; A268[13:6-10]. Notably, PacBio's argument in its reply brief did not even address claim 23 after ITRI pointed out these deficiencies in the prior art. *See generally* A409-46.

c. Facts relating to claims 27 and 28

Instead of using a mismatch to identify the locations of modified bases as claims 1-26 do, claims 27 and 28 require the incorporation of a "discriminating nucleotide analog" to determine modified base positions:

> [O]btaining sequencing data of the circular pair-locked molecule via single molecule sequencing, wherein at least one *nucleotide analog that discriminates between a base and its modified form* ["discriminating nucleotide analog"] is used to obtain sequence data comprising at least one position wherein the at least one differentially labeled nucleotide analog was incorporated; and . . . determining the positions of modified bases in the sequence of the double-stranded nucleic acid sample . . . .

A108[48:27-54] (emphasis added). ITRI's '630 patent states that sequence data can be "obtained using at least one nucleotide analog that discriminates between a base and its modified form (a 'discriminating analog'; it pairs preferentially with one but not the other of the base and its modified form)." A96[24:15-19]. Thus, the '630 patent defines a discriminating analog as an analog that pairs preferentially with one or the other of a base or a modified form of the base. *See supra* page 16.

The Board held that using a discriminating nucleotide analog as in claims 27 and 28 would have been obvious over the combination of the '375 patent in view of *Laird*. A42-3. But its analysis did not cite any disclosure of a discriminating nucleotide analog in either *Laird* or the '375 patent. *Id*.

Instead, the Board considered the act of "discriminating between methylated and non-methylated cytosines" sufficient. A42. As discussed above, however, *Laird* discriminated between methylated and unmethylated cytosines by comparing the same strand before and after treatment. *See supra* pages 30-31. *Laird* did not use a discriminating analog, and neither PacBio nor the Board asserted that it did.

The Board also did not address *Personal Genomes'* teaching to avoid using base-linked analogs. *See supra* page 18. Instead, the Board stated that "PacBio has not advanced an argument that the Personal Genomes Meeting Presentation is prior art if [PacBio's] Motion 1 [to rescind ITRI's benefit] is denied" and did not consider it further. A27 n.27. Although PacBios' Motion 1 was indeed denied, so

Personal Genomes was not 35 U.S.C. § 102(b) prior art, the Board ignored that *Personal Genomes* was still prior art under 35 U.S.C. § 102(a), based on undisputed declaration testimony that *Personal Genomes* was presented at a scientific meeting in October, 2008, before PacBios' December 12, 2008, priority date based on the '551 application and before ITRI's priority and filing dates in April and November, 2009. A115; A258; A744[¶¶ 2-3].

E. Facts Relevant to Whether the Prior Art Renders PacBio's Claims Unpatentable

The Board's rules provide for a presumption of cross-applicability such that, when "a motion for judgment of unpatentability against an opponent's claim on the basis of prior art is granted, each of the movant's claims corresponding to the same count as the opponent's claim *will be presumed to be unpatentable* in view of the same prior art *unless* the movant in its motion rebuts this presumption." 37 C.F.R. § 41.207(c) (emphasis added).

This rule cannot be waived and must be addressed in every interference. Specifically, when adopting 37 C.F.R. § 41.104(b), which allows for waiver or suspension of certain rules, the PTO stated that the presumptions in subpart E of the Rules, which includes Rule 207, "are outside the scope of § 41.104(b)." *Rules of Practice Before the Board of Patent Appeals and Interferences*, 49 Fed. Reg. 49960, 49983 (U.S.P.T.O. Aug. 12, 2004) ("*Interference Rules*").

PacBio asserted that its '375 patent qualifies as prior art under 35 U.S.C. § 102(e) (2006). A258. PacBio's counsel argued that the '375 patent was commonly owned with the invention at issue in the interference and therefore was not prior art to PacBio's claims under 35 U.S.C. § 103(c) (2006). A270. But ITRI contended this argument lacked any evidentiary support. A346[4-10].

Nevertheless, the Board permitted PacBio to attempt to cure this deficiency in the reply stage with a declaration from Stephen Moore, an employee at PacBio, alleging common ownership. A751. In an attempt to establish common ownership in that declaration, Mr. Moore submitted employee assignment agreements from less than half of the named inventors. Specifically, there is no evidence of an obligation to assign for eight named inventors of the '673 application (the PacBio application involved in the interference): Joseph Puglisi, Jessica Lee, Lei Jia, Jonas Korlach, Jon Sorenson, Dale Webster, John Lyle, and Jeremiah Hanes. *See* A482 (listing named inventors). ITRI informed the Board that at least one of them, Joseph Puglisi, apparently was a professor at Stanford University during the relevant time, approximately 2007-2009 (A460[10-12, n.5]).

Instead of providing evidence of assignment obligations from those eight named inventors, Mr. Moore stated that "[b]ased on information belief," they were not actually inventors of the Count. A752[¶6] (stating that "[b]ased on information and belief, the inventors of the subject matter of the count are" four of the initially

named inventors plus a new inventor, Cheryl Heiner); A482 (filing receipt for ’673 application); A251, section E (listing of ’178 application inventors). Mr. Moore provided no evidence or explanation justifying the omission of the named inventors beyond basing it on his “information and belief.” A752[¶6].

Before issuing its final decision, the Board acknowledged the presumption of cross-applicability, but concluded that inventorship was not important to the analysis, because the parties do not dispute that “PacBio *is* the assignee of the ’375 patent and ’673 and ’178 applications, no matter who the inventors *were* . . . .” A480 (emphasis added). “To that extent, *viz*., the issue of common ownership, the Board will consider the evidentiary weight of” the Declaration PacBio filed with its reply brief attempting to show common ownership of the ’375 patent and the involved applications “in support of PacBio’s arguments,” but will defer the issue of inventorship to the priority phase. A480. Thus, the Board indicated that the assessment of common ownership required only a determination that the applications were commonly assigned after the later application was filed. That assignment provided no information about the assignment obligations of each of the inventors at the proper point of time when common ownership was required─at the time of the invention as set forth in 35 U.S.C. § 103(c).

The Board’s final decision (A6-56) never explicitly addressed the presumption of cross-applicability, thus implicitly holding that the presumption of

cross-applicability did not invalidate PacBio's claims despite the lack of evidence of common ownership at the time of the invention.

VI. SUMMARY OF THE ARGUMENT

The Board erred in granting senior status to PacBio. The '551 application lacks a written description of the Count because it lacks any disclosure, express or inherent, of using a mismatch between the forward and reverse strand sequences of a CPLM to determine the position of the modified base. To the contrary, the '551 application expressly disclaims any "rel[iance] on the similarity of uracil to thymine." A920[¶23], which is exactly what the Count relies on when bisulfite treatment is involved. Furthermore, the single statement in paragraph 17 of the '551 application about comparing forward and reverse strands (A918-19[¶17]) cannot inherently support using a mismatch to identify a modified base position because it is undisputed that there are other methods that compare forward and reverse strands, including those described in the very applications PacBio incorporated by reference into that paragraph. Because the '551 application does not disclose an embodiment that falls within the Count, PacBio is not entitled to the benefit of its filing date.

The Board also erred in concluding that claims 1-26 were obvious because it engaged in hindsight. It relied on *Laird* for disclosure of determining a modified base position by comparing sequences of forward and reverse strands. *Laird*,

however, determined modified base positions in a strand by comparing modified and unmodified versions of the same strand, not the forward and reverse strands of a single molecule. Further, the Board's speculation that one skilled in the art *must* have recognized or understood that mismatches indicate modified base positions was unsupported by any prior art. The only place the Board could have learned that the position of a modified base can be determined by comparing forward and reverse strands is in ITRI's specification. The Board impermissibly used that teaching against the inventors.

The Board made an additional error in finding claim 23 obvious. Claim 23 comprises a step of accepting or rejecting reads of the sample sequence based on scores calculated for different known insert sequences positioned immediately upstream and/or downstream from the sample DNA. The Board never even alleged that the prior art taught this step.

The Board further erred in concluding that claims 27-28 would have been obvious because claims 27-28 require using a discriminating nucleotide analog, which was not disclosed by either of the two references the Board relied on (the '375 patent and *Laird*). The Board provided no explanation for why the missing element nonetheless would have been obvious. The Board also erred because it improperly disregarded the uncontroverted evidence of teaching away from the invention in PacBio's own *Personal Genomes* presentation.

Although the Board erred in finding that ITRI's claims were obvious, it compounded that error by failing to apply the presumption of cross-applicability and hold that PacBio's claims were unpatentable too. PacBio alleged that at the time the claimed inventions were made, they were commonly owned with the '375 patent. However, this bare allegation in the obviousness motion did not make the showing required by 37 C.F.R. § 41.207(c). On the legal question of inventorship, moreover, PacBio's late-filed declaration provided only a conclusory allegation made on information and belief. PacBio thus failed to show common ownership at the time of the invention or any other reason why its claims should not stand or fall with ITRI's, and the Board erred by not applying the prior art against PacBio's claims.

VII. ARGUMENT

A. Standard of Review

The Board's factual findings are reviewed for substantial evidence and its legal conclusions are reviewed de novo. *In re Baxter Int'l, Inc.*, 678 F.3d 1357, 1361 (Fed. Cir. 2012).

Written description is a question of fact. *Ariad Pharm., Inc. v. Eli Lilly & Co.*, 598 F.3d 1336, 1351 (Fed. Cir. 2010) (en banc). The proper interpretation of a legal standard, however, is a question of law. *See Romero v. United States*, 38 F.3d 1204, 1207 (Fed. Cir. 1994).

Obviousness is a question of law based on underlying factual findings. *Baxter Int'l, Inc.*, 678 F.3d at 1361. Four factors must be considered when determining whether an invention is obvious: (1) the scope and content of the prior art; (2) the differences between the prior art and the claims at issue; (3) the level of ordinary skill in the art; and (4) any relevant secondary considerations. *Graham v. John Deere Co. of Kansas City*, 383 U.S. 1, 17-18 (1966). A prior art reference must be considered in its entirety, i.e., as a whole, including portions that would lead away from the claimed invention. *W.L. Gore & Assoc., Inc. v. Garlock, Inc.*, 721 F.2d 1540, 1550 (Fed. Cir. 1983).

"Inventorship is a question of law, applied to relevant facts." *C.R. Bard, Inc. v. M3 Sys., Inc.*, 157 F.3d 1340, 1352 (Fed. Cir. 1998).

B. PacBio Is Not Entitled to Benefit of the '551 Application Filing Date Because the Application Does Not Disclose an Embodiment of the Count

To receive the benefit of an earlier application in an interference, the earlier application must contain a written description "for an *embodiment* within the count." *Hunt v. Treppschuh*, 523 F.2d 1386, 1389 (C.C.P.A. 1975). When "a party challenges written description support for an interference count or the copied claim in an interference, the originating disclosure provides the meaning of the pertinent claim language." *Agilent Techs., Inc. v. Affymetrix, Inc.*, 567 F.3d 1366,

1375 (Fed. Cir. 2009). The Count is ITRI's claim 24, and so it is construed in light of ITRI's '630 specification.

Written description "is not a question of whether one skilled in the art *might* be able to construct the patentee's [invention] from the teachings of the disclosure . . . ." *Martin v. Mayer*, 823 F.2d 500, 505 (Fed. Cir. 1987) (internal citation and quotation marks omitted). Rather, if claimed subject matter is not explicitly described, "the missing descriptive matter must *necessarily* be present in the . . . specification such that one skilled in the art would recognize such a disclosure." *Tronzo v. Biomet, Inc.*, 156 F.3d 1154, 1159 (Fed. Cir. 1998) (emphasis added).

The Board erred in two ways: 1) it found that the '551 application described using mismatches to locate modified bases (step (c)) without finding that it was *necessarily* present, as required for an inherent written description; and 2) its explanation of how the specification purportedly supports the limitation relied entirely on an example that the specification expressly disclaimed.

1. Inherent written description requires that the missing descriptive matter must *necessarily* be present, but disclosure of the '551 application does not necessarily disclose an embodiment of the Count

It is undisputed the '551 application does not contain an express disclosure of step (c) of the Count. The Board thus erred in finding what it considered to be the inherent disclosure of the Count in the '551 application where the disclosure was at best a possibility.

The only mention in the '551 application of comparing the reverse strand sequence *at all* is in a single sentence in paragraph 17:

> For example, sequence reads from the sense or 'forward' strand can be compared to sequence reads from the antisense or 'reverse' strand for the same nucleic acid template to further validate the existence of one or more modified bases in the template nucleic acid.

2026[¶17]; 2003[¶36]. ITRI submitted unrebutted evidence that this sentence disclosed the use of matches to confirm the reliability of sequencing because that is what is used in "redundant" and "consensus" sequencing (the only kind of sequencing disclosed in paragraph 17 and what Pacfic Biosciences was publicly promoting at that time). *See supra* page 24. Thus, the sentence could not have *necessarily* disclosed using mismatches to identify modified base positions. Indeed, the prior applications incorporated by reference into this paragraph indisputably teach *only* the use of matches to confirm reliability. *Id.*

Nevertheless, the Board found written description support for the Count. But in making this finding, it used the incorrect standard. Specifically, the Board found only that paragraph 17 did not *necessarily require* using matches to determine the position of modified bases. A54[18-21] ("[w]e do not agree with ITRI that the use of 'validate' in this passage [¶ 17] of the '551 application *necessarily requires* that the analysis of base pair [sic] only validate that the base pairs agree with the Crick-Watson rules.") (emphasis added). In other words, rather than considering whether

the ’551 application necessarily discloses using a mismatch to locate a modified base, which is required by the Count, the Board framed the issue as whether the ’551 application necessarily excluded the use of mismatches. In doing so the Board turned this Court’s law of inherency on its head.

Where a party, like PacBio here, must rely on inherency for a written description of the invention, it faces exacting scrutiny. As this Court’s predecessor famously stated, “[i]nherency . . . may not be established by probabilities or possibilities. The mere fact that a certain thing may result from a given set of circumstances is not sufficient.” *In re Oelrich*, 666 F.2d 578, 581 (C.C.P.A. 1981). Thus, the question was whether PacBio’s application *necessarily* disclosed using mismatches to locate the positions of modified bases. *Martin*, 823 F.2d at 505 (Written description “is a question [of] whether the application necessarily discloses that particular” invention.).

Paragraph 17 incorporates by reference several applications, all of which performed validation using matches and did not discuss using mismatches. *See supra* pages 16-20. Thus, when the ’551 application described verifying the accuracy of the sequencing results by comparing the sequences of the forward and reverse strands, it could not have excluded the known methods that were incorporated by reference that rely on matches to validate base locations

(confirming the reliability of the sequencing information).[6] Furthermore, validation of a *properly matched* base such as 5-methylcytosine, whose detection is a particularly preferred embodiment in the '551 application, A920[¶22], requires detecting *matches*. Paragraph 17 cannot be read to exclude validation of the positions of 5-methylcytosine. Given these undisputed facts, the Board could not conclude that the Count was *necessarily* disclosed in the '551 application, because it was just as likely, if not far more likely, that the application's sole reference to comparing forward and reverse strands concerned redundancy/consensus sequencing, which looks only at matches to confirm sequence reliability.

Rather than resolving what the '551 application *necessarily* conveyed, the Board engaged in conjecture about what one skilled in the art might have "recognized" and "understood." That a person in the art "would *expect to see* mismatched uracil-guanine base pairs" and "*understand* from that mismatch that" there was a modified base at that location cannot be more than a mere possibility. And that cannot satisfy this Court's stringent test for written description. *Martin*, 823 F.2d at 505.

---

[6] The '551 application's discussion of misincorporation rate does not remedy this gap in disclosure. As discussed *supra* at pages 20-21, using misincorporation rate is different from using mismatches in the sample, and misincorporation does not indicate whether there is a mismatch between the forward and reverse strands in the sample.

2. The ’551 application excludes relying on the similarity of uracil to thymine

One embodiment of the Count is determining the position of the modified base uracil by comparing sequence data of the forward and reverse strands. Importantly, when uracil is the modified base, the method of the Count depends on the *similarity* of uracil to thymine. *See supra* page 23. But the ’551 application expressly disclaims any method “relying on the similarity of uracil to thymine.” *Id*. That is, the ’551 application excludes any method that relies on that similarity, including the method of the Count when the modified base is uracil.

This Court’s case law requires considering the entire ’551 application, including paragraph 23, in evaluating the adequacy of a written description. *In re Edwards*, 568 F.2d. 1349, 1354 (C.C.P.A. 1978). But the Board never mentioned the ’551 application’s disclaimer. Instead, the Board expressly *relied* on the disclaimed similarity of uracil and thymine when it interpreted the one sentence in paragraph 17 that mentions comparing forward and reverse strands. A54[21]-55:2 (“[W]e find that *finding a mismatched uracil (or thymine)- guanine pair* would “validate”, commonly understood to mean “support or corroborate,” the *existence of a modified base* . . . .”). Thus, the Board erred by relying on finding a mismatched uracil to support the existence of a modified base because the ’551 application explicitly disclaimed it.

Since the ’551 application disclaims using the similarity of uracil and thymine to determine the position of a modified base, and does not describe any other mismatched modified base, the application lacks a description of an embodiment of the Count.

C. Claims 1-26 Would Not Have Been Obvious Because None of the References Taught Determining a Modified Base Position by Comparing Forward and Reverse Strand Sequences

The Board held that claims 1-26 would have been obvious over the ’375 patent and *Laird*. *See supra* page 34. The Board, however, engaged in impermissible hindsight in interpreting *Laird* to reach its conclusion of obviousness.

To properly evaluate obviousness, it “is difficult but necessary that the decisionmaker forget what he or she has been taught … about the claimed invention and cast the mind back to the time the invention was made (often as here many years), to occupy the mind of one skilled in the art who is presented only with the references, and who is normally guided by the then-accepted wisdom in the art.” *W.L. Gore*, 721 F.2d at 1553. Otherwise, one falls “victim to the insidious effect of a hindsight syndrome wherein that which only the inventor taught is used against its teacher.” *Id*.

1. The Board used impermissible hindsight to interpret *Laird*

Laird used the conventional method of distinguishing 5-methylcytosine from cytosine positions in bisulfite-treated DNA by comparing the sequence of each treated DNA strand to an untreated reference sequence of the same strand. *See supra* pages 30-31.

The Board improperly jumped to the conclusion that because mismatches are present in Figure 2, *Laird* used them to determine modified base positions. But *Laird* did not use these mismatches to identify modified base positions and did not suggest doing so. Indeed, no one reading *Laird* would consider using the mismatches that way because *Laird* had *already* identified the modified base positions in a different way. *See supra* pages 30-31. In fact, *Laird* never mentions the mismatches and the reader can barely see them with the naked eye. *See supra* pages 31-33.

Nonetheless, the Board stated that *Laird* sought to "examine the stability of cytosine methylation" by "comparing the forward and reverse strands of the DNA sample and identifying 'mismatched' base pair configurations." A39. This statement is wrong. *Laird* examined the stability of cytosine methylation by comparing positions of modified bases that had already been identified without using mismatches. A724; *see also supra* pages 30-32.

The Board also erred in stating that "*Laird* explicitly teaches looking for guanine-thymine mismatches as evidence of a non-methylated cytosine in a forward or reverse strand locus." A40. *Laird* never mentioned looking for mismatches, let alone looking for a mismatch to detect a modified base position. The Board leapt to the faulty conclusion about what *Laird* taught because its analysis was infected with hindsight. That "which only the inventors taught"—determining the position of a modified base by comparing forward and reverse strand sequences—was used against them. *W.L. Gore*, 721 F.2d at 1540.

The Board also noted *Laird*'s observation that "[b]isulfite conversion also reduces base pair complementarity . . ., a reduction that we found useful for PCR amplification, as well as for high-fidelity cloning and sequencing of hairpins containing unmethylated cytosines." A39 (quoting *Laird*, A723). There *Laird* pointed out that bisulfite conversion reduces base pair complementarity (i.e., causes mismatches). While *Laird* took advantage of this effect in PCR amplification, cloning, and sequencing, *Laird still used the traditional approach* of examining a reference sequence to determine which positions were or were not affected by bisulfite conversion. *Laird*'s failure to come up with the possibility of using mismatches to determine modified base pair positions actually supports the conclusion that doing so would not have been obvious.

So too does the failure of *Laird's* 2004 disclosure to lead any in the art to use or suggest mismatches for determining modified base positions before ITRI's invention. Particularly in light of the importance of modified bases in biology, "[i]f these discoveries and advances were routine and relatively easy, the record would undoubtedly have shown that some ordinary artisan would have achieved this invention within months of [the cited prior art]." *Leo Pharmaceutical Products, Ltd. v. Rea*, 726 F.3d 1346, 1354 (Fed. Cir. 2013). Instead of months, it took about five years after Laird's 2004 publication date before anyone (ITRI) discovered a method of identifying modified base positions by comparing forward and reverse strands.

In sum, the Board erred in reaching its legal conclusion of obviousness because it employed improper hindsight. Accordingly, the Board's determination that claims 1-26 would have been obvious should be reversed.

2. Claim 23 would not have been obvious because none of the references taught accepting or rejecting a sample sequence based on the score of an insert sequence

Contrary to the Board's statement that ITRI did not argue any claims separately (A31[n.29]), ITRI presented a separate argument for claim 23. A342. None of the asserted references disclosed, and PacBio did not even allege disclosure, of step (h) in claim 23.

"With respect to core factual findings in a determination of patentability, . . . the Board cannot simply reach conclusions based on its own understanding or experience—or on its assessment of what would be basic knowledge or common sense. Rather, the Board must point to some concrete evidence in the record in support of these findings." *In re Zurko*, 258 F.3d 1379, 1386 (Fed. Cir. 2001). "[A]ll factual differences which may be properly noted in any portion of a claim must be included within the basis for comparison with the prior art" in a proper obviousness analysis. *In re Duva*, 387 F.2d 402, 407 (C.C.P.A. 1967).

The Board failed to identify disclosure of step (h) of claim 23 in the prior art, contrary to *Zurko*. 258 F.3d at 1386. Because the Board did not do so, and the prior art did not disclose step (h), the decision that claim 23 would have been obvious should be reversed.

Claim 23 requires obtaining the score of an insert sequence adjacent to a sample sequence, and then using the insert score to accept or reject the sample sequence. *See supra* pages 34-36; A108[47:15-25]. In other words, the sample sequence is accepted or rejected based on the score of a *different* sequence, the insert. *Id.*

The Board's cursory obviousness analysis for claim 23 lacks any indication that the Board considered the element of using the score of a *different* sequence than the one being accepted or rejected in step (h). A41-42. Instead, it stated that

the ’375 patent “explicitly teaches determining a consensus sequence *at a position* from ‘multiple reads of *that position* of sequence.’” A41 (quoting the ’375 patent) (emphasis added). Thus, the Board applied a teaching of using reads of one position to determine the consensus at that *same* position, not different positions as claim 23 requires. The Board’s finding regarding Laird is similarly deficient. *See id.* (“*Laird* teaches determining sequences and agreement/disagreement of base pairs at given loci.”). Accordingly, the Board’s decision regarding claim 23 runs afoul of *Zurko* and *Duva*, and must be reversed.

D. Claims 27 and 28 Would Not Have Been Obvious

1. The Board erred because it failed to find a discriminating nucleotide analog in the applied prior art

Step (d) of claims 27 and step (c) of claim 28 recite obtaining sequence data using at least one nucleotide analog that discriminates between a base and its modified form. The ’630 patent defines a discriminating nucleotide analog as an analog that pairs preferentially with one or the other of a base or a modified form of the base. A96[24:15-19]. Thus, during sequencing, a discriminating nucleotide analog is present in addition to the nucleotides typically used (e.g., A, T, G, and C). And because that analog pairs preferentially with a base or its modified form, the presence of the discriminating nucleotide analog in the sequencing data indicates the position of the base or its modified form. *See supra* pages 36-37.

The Board held claims 27 and 28 obvious over the ’375 patent in view of *Laird*, but its decision failed to mention the use of discriminating nucleotide analogs in its discussion of the prior art. A42-43. Instead, it merely referred to the act of “discriminating between methylated and non-methylated cytosines” without addressing *how* discriminating was performed. A42. Thus, the Board failed to consider an element of claims 27 and 28─using a “nucleotide analog that discriminates between a base and its modified form” a distinct chemical reagent, when obtaining sequencing data. Neither the ’375 patent nor *Laird* disclosed any such analog or its use. A722-23 (only describing sequencing reactions as “conventional”); A260 (Pacific Bioscience’s admission that the ’375 patent did not explicitly disclose determining the position of any modified base at all, let alone doing so with a discriminating analog).

Much as with claim 23, the Board’s decision on claims 27 and 28 should be reversed because it runs afoul of *Zurko* and *Duva*. It did not consider the difference between the claims and the two references it relied upon, i.e., the requirement for using a discriminating nucleotide analog. *Duva*, 387 F.2d at 407. And it did not point to any concrete evidence in the record of why the claims were nonetheless obvious despite this difference from the cited prior art. *Zurko*, 258 F.3d at 1386.

2. The Board also erred by disregarding uncontroverted evidence that one of ordinary skill would have been led away from the invention

"[C]ompetent evidence tending to show the nonobviousness of appellants' invention . . . must be accorded fair weight" in the obviousness analysis. *In re Piasecki*, 745 F.2d 1468, 1474 (Fed. Cir. 1984). Proceeding "contrary to the accepted wisdom" is "strong evidence of unobviousness." *In re Hedges*, 783 F.2d 1038, 1041 (Fed. Cir. 1986) (internal citation and quotation marks omitted). PacBio asserted the *Personal Genomes* presentation as prior art under 35 U.S.C. § 102(a) based on evidence that it was presented at a scientific meeting in September 2009. A258; A744[¶¶2-3]. The status of *Personal Genomes* as prior art was undisputed. A344[9-11]. The Board nonetheless refused to consider *Personal Genomes* based on the mistaken understanding that it had not been established as prior art at all because it was not § 102(b) art. A27[n.27]. There is no basis for the Board's refusal to consider *Personal Genomes*.

Personal Genomes would have led one of ordinary skill away from the invention because it taught that base-linked nucleotide analogs are problematic. A946-47; A1022-23[¶69]. PacBio did not introduce any evidence to counter the teaching away in its own *Personal Genomes* presentation. A443[¶136] (PacBio's admission that its sole expert did not opine on obviousness). And the prior art did not disclose discriminating nucleotide analogs that are not base-linked nucleotide

analogs. A1022-23[¶¶68, 71]. Therefore, one of ordinary skill would not have done what PacBio's *Personal Genomes* presentation said *not to do* and modified the '375 patent to use base-linked analogs in single molecule sequencing.

E. The References Asserted by PacBio Must Be Applied Against Its Own Claims Because It Did Not Rebut the Presumption of Cross-Applicability

Under 37 C.F.R. § 41.207(c), when "a motion for judgment of unpatentability against an opponent's claim on the basis of prior art is granted, each of the movant's claims corresponding to the same count as the opponent's claim *will be presumed to be unpatentable in view of the same prior art unless the movant in its motion rebuts this presumption*." (Emphasis added.)

Although this issue becomes moot if the Court holds that none of ITRI's claims would have been obvious, PacBio did not rebut the presumption of obviousness of its own claims with evidence, and the Board has failed to properly apply the presumption. If the Court affirms the Board's holding of obviousness as to any of ITRI's claims, then the Court should hold PacBio's claims to be invalid for the same reason in view of PacBio's failure to rebut the presumption. 37 C.F.R. § 41.207(c). PacBio did not provide timely evidence to rebut the presumption in its obviousness motion, and its late-filed purported evidence was insufficient.

1. The Board misapplied 35 U.S.C. § 103(c) and implicitly determined that the presumption did not apply to PacBio

PacBio asserted the ’375 patent as 35 U.S.C. § 102(e) prior art but alleged that it was disqualified as prior art against PacBio’s own claims under 35 U.S.C. § 103(c). A270. PacBio did not allege that *Laird* or the other secondary references were not prior art against its own claims.

The Board failed to explain why the presumption of unpatentability did not apply to PacBio even though that issue was in dispute. A346. Unless the Board simply forgot to address the issue, one can assume that the Board implicitly determined that the presumption did not apply. But there was insufficient evidence to reach that conclusion and the Board appeared to misunderstand what was required to disqualify prior art under 35 U.S.C. § 103(c).

Under 35 U.S.C. § 103(c), 35 U.S.C. § 102(e) prior art such as the ’375 patent does not apply if “the subject matter and the claimed invention were, *at the time the claimed invention was made*, owned by the same person or subject to an obligation of assignment to the same person.” (Emphasis added.) The Board misapplied § 103(c) because it concluded that inventorship was not important to the analysis, indicating that the assessment of common ownership required only a determination that the applications were commonly assigned after the later application was filed. A480 (“That PacBio *is* the assignee of the ’375 patent and

’673 and ’178 applications, no matter who the inventors *were*, is not disputed by the parties.”) (emphasis added). This was error because common ownership of the applications after they were filed does not establish common ownership or any obligation to assign ownership at the time of the invention, which here was on or before PacBio’s first non-provisional filing date, December 10, 2009. A115.

Whether there was common ownership at the time of the invention (on or before December 10, 2009) depends on who the inventors were and whether they had obligations to assign the invention to PacBio. *See Beech Aircraft Corp. v. EDO Corp.*, 990 F.2d 1237, 1248 (Fed. Cir. 1993) (“At the heart of any ownership analysis lies the question of who first invented the subject matter at issue, because the patent right initially vests in the inventor . . . .”).

The Board deferred the issue of inventorship to the priority phase. A480. Thus, the Board never conducted the proper analysis of whether there was common ownership at the time of PacBio’s claimed invention. But the presumption of cross-applicability is non-waivable. *See supra* page 38. Thus, the presumption of cross-applicability should have attached, and the Board should have applied the prior art against PacBio’s claims if any claim of ITRI’s was found unpatentable over the prior art.

2. PacBio was required to prove common ownership at the time of the claimed invention in its motion, but did not, and the presumption of cross-applicability should have attached

PacBio had the burden of proof on common ownership at the time of its invention. § 41.207(c). The Board's rules require evidence to support motions. 37 C.F.R. § 41.208(b) (A "motion must provide a showing, supported with appropriate evidence . . . ."). Other rules and the Board's Standing Order further confirm that "evidence" and "portions of the record that support [a material] fact" must be provided. 37 C.F.R. §§ 41.121(c)(1)(ii), (d)(1); A155[¶121.3.3] (Board's Standing Order). Furthermore, § 41.207(c) itself speaks of a "presumption" that must be rebutted, and presumptions are rebutted with evidence, not attorney argument. Furthermore, PacBio had to rebut the presumption of cross-applicability "in its motion" or suffer the loss of its corresponding claims. *Id*.

Attorney argument is not evidence. *In re Shulze*, 346 F.2d 600, 602 (C.C.P.A. 1965). PacBio's motion provided attorney argument, but no proof. A270.

3. PacBio's reply declaration could not show common ownership because it failed to establish a different inventorship

Following PacBio's failure to rebut the presumption in its motion, it violated the rules against introducing new issues in its reply brief. 37 C.F.R. § 41.122(b). As the Board acknowledged, PacBio asserted a new inventorship of its involved

applications at the reply stage, when ITRI had no opportunity to respond. A478-79. Even worse, the declaration PacBio filed alleged a dramatically different inventorship "on information and belief" that dropped eight inventors listed in the involved applications and added one. *See supra* pages 39-40.

Unsupported, conclusory allegations are not evidence. *See, e.g.*, *Paragon Podiatry Lab., Inc. v. KLM Labs., Inc.*, 984 F.2d 1182, 1188 (Fed. Cir. 1993) (disregarding "conclusory legal opinion" on legal issue of on-sale bar versus experimental use). Here, there are no facts supporting PacBio's conclusory allegation on the legal issue of inventorship. Furthermore, an unsupported allegation on information and belief is only a statement that the declarant believes the asserted matter. In other words, PacBio came forward only with what one of its employees thought about the legal issue of inventorship, without any explanation of the basis. This is not evidence of inventorship. *See HollyAnne Corp. v. TFT, Inc.*, 199 F.3d 1304, 1309 (Fed. Cir. 1999) (disregarding unsupported recitation in affidavit "on information and belief").

Because PacBio never submitted any factual evidence that support its asserted inventorship, the original inventorship of PacBio's applications should apply to its ownership status under 35 U.S.C. § 103(c). *Brown v. Edeler*, 110 F.2d 858, 861 (C.C.P.A. 1940) ("an application . . . made by two or more persons claiming to be joint inventors, is prima facie evidence that they are such"). PacBio

never provided evidence that the eight named inventors its declarant believed were not inventors had obligations to assign the invention to PacBio and one of them was apparently a professor at Stanford University during the relevant time. A460[10-12, n.5]. Accordingly, PacBio failed to prove common ownership at the time of the invention, and if the Court affirms the Board's holding of obviousness as to a claim of ITRI's, then the Court should hold all of PacBio's claims invalid for the same reason.

VIII. CONCLUSION

For the foregoing reasons, the Court should reverse the Board's determinations that PacBio was entitled to the benefit of the '551 application and that '630 patent claims 1-28 would have been obvious. However, if any of the '630 patent claims remain invalid as obvious, the prior art references should be applied against the claims of PacBio's involved applications.

Respectfully submitted,

March 19, 2015 */s/ David S. Forman*

David S. Forman
FINNEGAN, HENDERSON, FARABOW, GARRETT & DUNNER, LLP
901 New York Avenue, NW
Washington, DC 20005-4413
(202) 408-4000

Erik R. Puknys
Adam M. Breier
FINNEGAN, HENDERSON, FARABOW, GARRETT & DUNNER, LLP
3300 Hillview Avenue
Palo Alto, CA 94304-1203
(650) 849-6600

Attorneys for Appellant

ADDENDUM

BoxInterferences@uspto.gov
Telephone: 571-272-4683 Entered: August 3, 2014

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

Patent Interference No. 105,970 (JGN)

INDUSTRIAL TECHNOLOGY RESEARCH INSTITUTE
and Ti-Shiue Biotech, Inc.
Junior Party
(US 8,846,630)

v.

PACIFIC BIOSCIENCES
of California, Inc.
Senior Party
(13/633,673 and 13/930,178)

JUDGMENT
Bd.R. 127

Before SALLY GARDNER LANE, JAMES T. MOORE, and JOHN G. NEW, *Administrative Patent Judges.*

NEW, Administrative Patent Judge

The Board determined that all of the claims of Junior Party Industrial Technology Research Institute's ("ITRI") US Patent No. 8,486,630 B2 are unpatentable under 35 U.S.C. § 103(a).[1] Accordingly there is no basis to continue the interference. It is—

ORDERED that judgment be entered against Junior Party ITRI for count 1[2];

FURTHER ORDERED that claims 1-28 of ITRI's involved U.S. Patent No. 8,486,630 B2 be CANCELED, 35 U.S.C. 135(a)[3]; and

FURTHER ORDERED that a copy of this judgment be entered in the administrative records of the involved 8,486,630 B2 patent and application.

NOTICE: "Any agreement or understanding between parties to an interference, including any collateral agreements referred to therein, made in connection with or in contemplation of the termination of the interference, shall be in writing and a true copy thereof filed in the Patent and Trademark Office before the termination of the interference as between the said parties to the agreement or understanding." 35 U.S.C. 135(c); see also Bd.R. 205 (settlement agreements).

---

[1] Paper No. 167
[2] Paper No. 1
[3] As was in effect on March 15, 2013. *See* Pub. L. 112-29, § 3(n), 125 Stat. 284, 293 (2011).

Attorney for ITRI:

David S. Forman, Ph.D.
M. Paul Barker
Adam M. Breier, Ph.D.
Finnegan, Henderson, Farabow, Garrett & Dunner, LLP
David.forman@finnegan.com
Paul.barker@finnegan.com
Adam.breier@finnegan.com

Attorney for PacBio:

Michael S. Tuscan, Ph.D.
Bonnie Weiss McLeod, Ph.D.
Cooley, LLP
mtuscan@cooley.com
bweissmcleod@cooley.com
zpatdocketing@cooley.com

BoxInterferences@uspto.gov
Telephone: 571-272-4683 Entered: October 2, 2014

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

Patent Interference No. 105,970 (JGN)

INDUSTRIAL TECHNOLOGY RESEARCH INSTITUTE
and Ti-Shiue Biotech, Inc.
Junior Party
(US 8,846,630)

v.

PACIFIC BIOSCIENCES
of California, Inc.
Senior Party
(13/633,673 and 13/930,178)

ERRATA

The Judgment, entered on September 3, 2014 contained an incorrect date in the heading. The date was listed as August 3, 2014. It should have been listed as September 3, 2014.

/Amy Kattula/
Paralegal Specialist

Attorney for ITRI:

David S. Forman, Ph.D.
M. Paul Barker
Adam M. Breier, Ph.D.
Finnegan, Henderson, Farabow, Garrett & Dunner, LLP
David.forman@finnegan.com
Paul.barker@finnegan.com
Adam.breier@finnegan.com

Attorney for PacBio:

Michael S. Tuscan, Ph.D.
Bonnie Weiss McLeod, Ph.D.
Cooley, LLP
mtuscan@cooley.com
bweissmcleod@cooley.com
zpatdocketing@cooley.com

BoxInterferences@uspto.gov
Telephone: 571-272-4683 Entered: September 3, 2014

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

Patent Interference No. 105,970 (JGN)

INDUSTRIAL TECHNOLOGY RESEARCH INSTITUTE
and Ti-Shiue Biotech, Inc.
Junior Party
(US 8,846,630)

v.

PACIFIC BIOSCIENCES
of California, Inc.
Senior Party
(13/633,673 and 13/930,178)

Before SALLY GARDNER LANE, JAMES T. MOORE, and JOHN G. NEW, *Administrative Patent Judges*.

NEW, *Administrative Patent Judge*.

DECISION ON MOTIONS
37 C.F.R. § 41.125

I. Introduction

A. Background

Interference 105,970 is before a panel of the Board for a decision on motions. Oral argument was requested.[1]

We have considered the record submitted by the parties.

B. Abbreviations

The following **abbreviations** are used in the record and this opinion.

Pacific Biosciences Motion 1 (PacBio Motion 1)

Pacific Biosciences Motion 2 (PacBio Motion 2)

Industrial Technology Research Institute Motion 1 (ITRI Motion 1)

Industrial Technology Research Institute Motion 2 (ITRI Motion 2)

Industrial Technology Research Institute Motion 3 (ITRI Motion 3)

Industrial Technology Research Institute Opposition 1 (ITRI Opp. 1)

Industrial Technology Research Institute Opposition 2 (ITRI Opp. 2)

Pacific Biosciences Opposition 1 (PacBio Opposition)

C. Parties

1. Junior Party

The Junior Party is Pacific BioSciences of California, Inc. ("PacBio"), which is also the real party-in-interest.[2]

2. Senior Party

The Senior Party is Industrial Technology Research Institute

[1] Paper Nos. 155, 156

[2] Paper No. 5

and Ti-Shiue Biotech, Inc. ("ITRI"), which are also the real parties-in-interest.[3]

D. Subject matter involved in interferences

1. The counts

Count 1 is the sole count in this interference.[4]

2. Count 1

ITRI claim 24:

A method of determining a sequence of a double-stranded nucleic acid sample and a position of at least one modified base in the sequence, comprising:

a. locking the forward and reverse strands of the nucleic acid sample together to form a circular pair-locked molecule;

b. obtaining sequence data of the circular pair-locked molecule via single molecule sequencing, wherein sequence data comprises sequences of the forward and reverse strands of the circular pair-locked molecule; and

c. determining the sequence of the double stranded nucleic acid sample and the position of the at least one modified base in the sequence of the double stranded nucleic acid sample by comparing the sequences of the forward and reverse strands of the circular pair-locked molecule, wherein at least one modified base in the double stranded nucleic sample is paired with a base having a base pairing specificity different from its preferred partner base.

The claims of the parties are:

[3] Paper No. 9, 17

[4] Papers Nos. 1, 28

ITRI:	Claim 1-28; all corresponding to Count 1
PacBio:	13/633,673 application; claim 1 corresponding to Count 1 13/930,178 application: claims 1-4, all corresponding to Count 1[5]

With respect to Count 1, the parties have been accorded the benefit of the following applications:

ITRI:	Provisional application 61/167,313, filed **7 April 2009**
PacBio:	Application 12/635,618, filed 10 December 2009 Provisional application 61/201,551, filed **11 December 2008**

E. Subject matter involved

Knowing the sequence of the nucleotide base pairs in a molecule of nucleic acid, such as DNA, provides the fundamental key to understanding genomic structure and function. A molecule of DNA consists of two strands (5′-3′ and 3′-5′ or "forward" and "reverse" strands) of paired nucleotide bases, the bases being paired according to the Crick-Watson nucleotide base-pairing rules, i.e., adenine–thymine and cytosine–guanine. The forward and reverse strands therefore usually contain complementary and redundant information, meaning that the sequence of one strand can be deduced from the other using the base pairing rules.

Chemical modifications to the nucleotide bases can result from exposure to damaging agents such as radiation or reactive oxygen species and can have epigenetic or mutagenic effects. ITRI Motion 2[6] at 3. It is therefore often important to know not only the base sequence of a nucleic acid, but also the positions of modified bases. When

[5] *See* Paper No. 28 at 3.
[6] Paper No 59

Pacific Biosciences' Appl. No. 61/201,551 (the "'551 application") was filed, sequencing technology could be used to obtain genetic sequence data from single nucleic acid molecules using a sequencing-by-synthesis approach—a type of single molecule sequencing. *Id.* at 3-4. Sequencing-by-synthesis involves determining the order in which nucleotides are added during complementary nucleic acid strand synthesis, and then determining the identity of the nucleotides in the original molecule using the usual base-pairing rules. *Id.* at 4. However, the results of sequencing-by-synthesis of a single DNA strand based on the Crick-Watson rules do not identify whether or where chemical modifications of the bases are present. *Id.*

Count 1 of this interference is claim 24 of ITRI's US 8,846,630 B2[7] (the "'630 patent") (Ex. 2001) and is directed to detecting a mismatch in a constructed circular pair-locked molecule (CPLM); a single circular molecule that contains within it both the forward and reverse strands that are to be analyzed. Motion 1 at 4. CPLMs are formed by ligating "hairpin" sequences to each end of the DNA segment to be analyzed, one hairpin segment containing a primer sequence to initiate replication by DNA polymerase. Performing sequencing-by-synthesis on a CPLM can then generate sequence data for both of the original forward and reverse complementary strands. *Id.* Disagreements in the analyzed sequences of the forward and reverse strands indicate that the CPLM has a mismatched, modified base in at least one position; an investigator can thus use the disagreement to determine the position of the modified base. *Id.* In Count 1, one compares the sequences of the forward and reverse strands, and determines that a modified base is present at positions where there is a disagreement. *Id.* at 5.

[7] Paper No. 62

II. PacBio's Motion 1[8] to rescind the benefit accorded ITRI for the '3l3 application

PacBio moves to deny the benefit accorded to ITRI for its Provisional Application 61/167,313 (the "'3l3 application") because it allegedly fails to provide a constructive reduction to practice of Count 1. Motion at 1. According to PacBio, ITRI's '313 application fails to: (1) adequately provide a written description of an embodiment within the scope of Count 1; and (2) enable an embodiment within the scope count 1. Motion at 2, 9.

A. Written Description

PacBio argues that ITRI's '313 application fails to provide an adequate written description of two of the limitations of Count 1: (1) "locking the forward and reverse strands of the nucleic acid sample together to form a circular pair-locked molecule"; and (2) "obtaining sequence data of the circular pair-locked molecule by single molecule sequencing." Motion at 3, 8.

1. "locking the forward and reverse strands of the nucleic acid sample together to form a circular pair-locked molecule"

PacBio argues that the '313 application fails to describe the length of the double stranded DNA sample that can be used in the manufacture of a CPLM. Motion at 3-4 (citing Ex. 1003[9]). Nor, argues PacBio, does the '313 application provide a description of the nucleotide length, base composition or the strandedness (e.g., double stranded, single stranded or a combination) of the remaining components necessary to construct a

[8] Paper No. 55
[9] Paper No. 38

CPLM. *Id.* PacBio asserts that the series of steps required for constructing a CPLM and the associated reaction conditions are similarly not provided in the '313 application. *Id.* PacBio contents that one of skill in the art would therefore be left to determine the nature of the components, series of steps and reaction conditions necessary to build the CPLM sequencing template: this, contends PacBio, does not meet the standard for achieving a constructive reductive to practice. *Id.* (citing *Goeddel v. Sugano*, 617 F.3d 1350, 1355-1356 (Fed. Cir. 2010).

Furthermore, argues PacBio, Figures 1A and Figure 3A of the '313 application depict CPLMs that include gaps between the hairpin endpieces and the central double stranded region such that the forward and reverse strands are not completely locked together, and that no examples, prophetic or otherwise, are provided with respect to how the forward and reverse strands of the nucleic acid sample are locked together. Motion at 4-5 (citing Ex. 1003). PacBio admits that although ligation reactions and hairpin DNA sequences were publically available as of the '313 application filing date, these features were not mentioned in the '313 application as useful for manufacturing a CPLM or incorporated by reference in the application. Motion at 5 (citing Ex. 1003).

PacBio argues further that although the '313 application recites "[t]he invention discloses a method for preparing a unique DNA template by bridging the two complementary strands of a DNA fragment together," the application fails to describe which steps must be performed to bridge the strands together to form the CPLM. Motion at 5-6 (citing Ex. 1003).

PacBio adduces the Declaration of Dr. Gerald Zon[10] (the "Zon Declaration 1") (Ex. 1010). Dr. Zon holds a doctoral degree in organic chemistry from Princeton University, has had a career of over 40 years in academia and the private sector, and has authored over 270 publications. *Id.* Dr. Zon is currently Director of Business Development for TriLink BioTechnologies. *Id.* After reviewing his curriculum vitae, we find that Dr. Zon is an expert qualified to opine about the issues before the Board in this interference.

Dr. Zon states that there are two classes of methods for constructing a CPLM: the first is to create single stranded overhangs on a double stranded DNA fragment, followed by hybridization of the double stranded DNA fragment to hairpin DNAs also having single stranded overhangs, followed by gap filling and ligation to form the fully circularized molecule. Motion at 6, (citing Ex. 1010, ¶¶14-15). The second is to create a double stranded DNA fragment with blunt ends followed by blunt end ligation of hairpin DNA to each side to circularize the molecule. *Id.* Zon declares that, with respect to the first method, although the arrows in Figures 1A and 3A appear to depict a gap filling step, there is no overhang shown to perform a hybridization step and that it is unclear how one would fill the gaps. *Id.* (citing Ex. 1010, ¶ 16).

With respect to the second method, Zon declares that the arrows in Figures 1A and 3A of the '313 application are not consistent with a blunt end ligation approach. *Id.* (citing Ex. 1010, ¶ 17).

PacBio admits that, at the time the '313 application was filed, PacBio's own disclosure of CPLM sequencing templates had been publically available for

[10] Paper No. 96

approximately six months. Motion at 9 (Ex. 1009[11] at 15-22). However, argues PacBio, Ex. 1009, which discloses how to assemble CPLMs for use as sequencing templates, was neither incorporated-by-reference nor cited in the '313 application. *Id.*

ITRI responds that PacBio has adduced no testimonial evidence with respect to what the '313 application would have conveyed to one of ordinary skill in the art, and that the remainder of its arguments constitute mere attorney argument by those not qualified to so opine. Opp. at 4. According to ITRI, in Zon Declaration 1 and also in Zon Declaration 2[12] (Ex. 1020), Dr. Zon discusses Figures, 1A and 3A, but does not discuss any of the text or the remaining figures in the application. *Id.* at 5 (citing Ex. 1010, ¶¶ 1-19; 1020, ¶¶ 1-21). Moreover, argues ITRA, with respect to Figures 1A and 3A, Dr. Zon says nothing about what those figures would have conveyed to one skilled in the art. *Id.* Furthermore, ITRI argues, even if, *arguendo*, the Figures were incompatible with making a CPLM, it would not meet PacBio's burden of proof—ITRI argues that PacBio must analyze the disclosure within the "four corners of the specification" in light of the state of the art. *Id.* (quoting *Ariad v. Lilly*, 598 F.3d 1336, 1351 (Fed. Cir. 2010) (en banc)).

Consequently, argues ITRI, Dr. Zon's testimony is fatally flawed because he does not state whether the '313 application would have conveyed possession of an embodiment of the Count to one skilled in the art, despite declaring that he believed himself to be able to comment on the level of ordinary skill in the art. *Id.* (citing Ex. 1010, ¶ 8).

[11] Paper No. 44
[12] Paper No. 103

Furthermore, argues ITRI, Dr. Zon declares that two methods involving hairpin ligation could have been used to make a CPLM. *Id.* (citing Ex. 1010, ¶¶ 14-15). Moreover, ITRI argues PacBio admits that hairpins and ligations were known. *Id.* (citing Motion at 5, ll.11-12). Consequently, argues ITRI, Dr. Zon's testimony, read together with the '313 application's teaching to make a CPLM from target DNA, supports the conclusion that the application would have conveyed possession of making a CPLM to one of ordinary skill. *Id.*

Similarly, ITRI argues that PacBio provides no other evidence supporting its argument that the '313 application would not have conveyed, to one of ordinary skill in the art, obtaining sequence data of a CPLM via single molecule sequencing. Opp. at 5. ITRI points out that PacBio admits that "single molecule sequencing" appears in the '313 application, and admits that methods of single molecule sequencing of a CPLM were known. *Id.* (citing Motion at 8, ll. 15-16; 10, ll. 9-12; Ex. 1009 at 15-22). ITRI asserts that a single mention of a term suffices for written description of well-known subject matter. *Id.* (citing *EnOcean GmbH v. Face Int'l Corp.*, 742 F.3d 955, 961 (Fed. Cir. 2014). ITRI contends that PacBio's pointing out that a term appears once cannot, by itself, establish a lack of description. Opp. at 6.

ITRI maintains that the '313 application describes and illustrates target DNA from which the CPLM is made and that that a CPLM is made by connecting the strands of the target DNA, and it showed that linkers are used to do so. Opp. at 6-7 (citing Ex. 2002 7 at 5, Fig. 2, ¶¶ 1-2; Ex. 2034, ¶¶ 35-37). It explained "that a CPLM is made by connecting the strands of the target DNA, and it showed that linkers are used to do so." Opp. at 7. (Ex. 2002, p. 5, paragraphs 3-4; p. 6, Fig. 4A.1; p. 7, Fig. A2; Ex. 2034, ¶ 37). ITRI contends that a person of ordinary skill would have understood from the '313

application that the inventors described a circular DNA molecule having a double-stranded "target DNA" region bracketed by two linkers that connect the two strands and that one linker in the molecule contains a binding site for a sequencing primer. Opp. at 7 (citing Ex. 2034, ¶¶ 37, 42).

In support of this contention, ITRI adduces the declaration of Dr. Shawn Levy (Ex. 2034, the "Levy declaration"). Dr. Levy received his doctoral degree in Biochemistry and Genetics from Emory University in 2000 and is currently a Faculty Investigator and the Director of the Genomic Services Laboratory at the HudsonAlpha Institute for Biotechnology in Huntsville, Alabama. *Id.* at ¶¶ 1, 4. Dr. Levy has published more than 90 peer-reviewed articles on the subject of the application of genomic technologies. *Id.* at ¶ 6. Having reviewed Dr. Levy's Declaration we find that he is an expert qualified to opine on the issues before the Board in the instant interference.

Dr. Levy states that one of ordinary skill in the art would have a Ph.D. in molecular biology or a related field and two years' experience in a research laboratory where sequencing is commonly performed. Ex. 2034, ¶ 15. Dr. Levy declares that one so skilled in the art would have known of at least three ways to attach the hairpin linkers, each involving ligation and that such ligation reactions were well known in the art as of April 2009. Ex. 2034, ¶ 51.

Consequently, Dr. Levy attests, the design of ligation reactions that would convert a target DNA to a CPLM, e.g., by attaching hairpin sequences would have been within the level of ordinary skill as of April 7, 2009, given the guidance in the '313 application Ex. 2034, ¶¶ 19-20 (citing Ex. 2002 at Fig. 2, Fig. 4A.1 and accompanying text). Dr. Levy also declares that the outcome of these ligation reactions would have been

predictable to one skilled in the art based on both general understanding of ligation reactions and specific, known examples of hairpin ligations. *Id.*

ITRI also adduces Exhibits 2037, 1009, and 2038 as examples demonstrating that ligation techniques were well known and predictable in the art at the time the '313 application was filed.

2. "obtaining sequence data of the circular pair-locked molecule by single molecule sequencing"

PacBio concedes that the '313 application discusses the use of "single molecule sequencing" but argues that it includes only a single mention that the technique can be used in conjunction with a pair locked molecule. Motion at 9 (citing Ex. 1003 at 3). According to PacBio, the '313 application is silent regarding the steps necessary to carry out a single molecule sequencing reaction on a CPLM, the types of sequencing reactions that are amenable for use with a CPLM, and whether the CPLM is assembled in a different manner based on the sequencing reaction employed. Motion at 9 (citing Ex. 1003).

ITRI responds that single molecule sequencing techniques were well known and points to PacBio's US 7,302,146, (Ex. 2036[13], the "'146 patent") issued in 2007, that discloses single molecule sequencing of circular templates. Opp. at 11 (citing, e.g., Exs. 2036; 2034, ¶¶ 45-46). Dr. Levy declares that the '146 patent teaches that its "sequencing methods can be used" with "any nucleic acid molecule, including double-stranded or single-stranded, linear or circular nucleic acids (e.g., circular DNA), single stranded DNA hairpins," and various other structures including "complex nucleic acid

[13] Paper No. 113

structures" such as "tandem repeats." *Id.* (citing Ex. 2034, ¶ 45; Ex. 2036, col. 23, ll. 50-58). Therefore, Dr. Levy declares, one skilled in the art would have known of a single molecule sequencing method suitable for sequencing a circular nucleic acid molecule, such as a circular pair-locked nucleic acid molecule. *Id.* (citing Ex. 2034, ¶ 45). Furthermore, there would have been no concern that double-stranded regions or hairpin linkers contained within a CPLM might present difficulties because the '146 patent expressly indicates that its methods work with all of these types of structures. *Id.*

3. The '313 application conveyed possession of a single embodiment within count 1

We are not persuaded by either of PacBio's arguments.

As an initial matter, Dr. Zon opines that, as of April 7, 2009, the ordinarily skilled artisan had an advanced degree (i.e., M.S. or Ph.D. degree) in molecular biology, biochemistry, biophysics, bioengineering, or related field, and one or more years post-graduate experience. Ex. 1010, ¶ 8. Similarly, Dr. Levy opines that, at approximately the same period, one of ordinary skill in the art of single molecule sequencing and detection of modified bases would have a Ph.D. in molecular biology or a related field and two years' experience in a research laboratory where sequencing is commonly performed. Ex. 1034, ¶ 15. We agree with the parties that a person of ordinary skill in the art would have possessed a Ph.D. in molecular biology or a related field and 1-2 years' postdoctoral experience in a position where sequencing is routinely performed. Having set the standard of ordinary skill, therefore, at a relatively high level, we assume that a person of ordinary skill would have a sophisticated understanding of the molecular biology of nucleic acids and their synthesis.

Benefit of an earlier-filed application in an interference requires that the application would have conveyed possession of a single embodiment within the Count to one of ordinary skill in the art. Bd.R. 201; *see also Hunt v. Treppschuh*, 523 F.2d 1386, 1389 (C.C.P.A. 1975). The only substantive testimonial evidence that PacBio adduces in support of its arguments is the Zon Declarations. Dr. Zon declares that:

> Based on my understanding of the invention claimed in the '630 patent and the nature of a circular pair-locked molecule (cPLM), I envision two classes of methodology for constructing a cPLM.
>
> The first is to create single stranded overhangs on a double stranded DNA fragment, followed by hybridization of the double stranded DNA fragment to hairpin DNAs also having single stranded overhangs, followed by gap filling and ligation to form the fully circularized molecule. The second is to create a double stranded DNA fragment with blunt ends followed by blunt end ligation of hairpin DNA to each side to circularize the molecule.
>
> With respect to the first method, although the arrows in Figures 1A and 3A appear to depict a gap filling step, there is no overhang shown to perform a hybridization step. Additionally, it is unclear how one would fill the gaps identified by the red circles in Figures 1 A and 3A.
>
> With respect to the second method, the arrows in Figures 1A and 3A are not consistent with a blunt end ligation approach.
>
> Accordingly, Figures 1A and 3A are not compatible with the steps and/or components necessary to construct a cPLM.

Ex. 1010, ¶¶ 14-18. The entire relevant portion of the Zon Declaration, therefore, deals with whether, in Dr. Zon's opinion, Figures 1A and 3A are "compatible with the steps and/or components necessary to construct a CPLM." Ex. 1010, ¶ 18.

However, Zon does not discuss the state of the prior art, other than to observe that it was well known in the art to store a double-stranded DNA sequencing template at or below 25°C and in the absence of divalent cations. Ex. 1010, ¶¶ 11, 12. Dr. Zon is silent, however, with respect to the known state of the art of ligation and the construction of CPLMs at the time the '313 application was filed. Nor does Dr. Zon declare what a person of ordinary skill at the time of invention would have understood of the construction of CPLMs or whether such a skilled artisan would be in possession of the invention. And Dr. Zon is also silent with respect to the subject of sequencing methods, single molecule or otherwise.

By contrast, ITRI's declarant, Dr. Levy, states that:

> [O]ne skilled in the art would have known of at least three ways to attach the hairpin linkers. Each of these involved ligation. Ligation reactions were well known in the art. They had been known and performed for decades as of April 2009. And ligations were a common, basic step in molecular biology cloning techniques … Furthermore, the outcome of these ligation reactions would have been predictable to one skilled in the art based on both general understanding of ligation reactions and specific, known examples of hairpin ligations.
>
> Thus, there was knowledge in the art of relevant procedures for making a CPLM, and these procedures were predictable. One of ordinary skill would have understood which procedures would have been suitable given the guidance in the '313 application. Accordingly, there would have been little, if any, need for one of ordinary skill in the art to experiment to lock the forward and reverse strands of a nucleic acid sample together to form a circular pair-locked molecule.

Ex. 2034, ¶¶ 51, 52 (internal citations omitted). Moreover, Dr. Levy declares that:

> Techniques for single molecule sequencing of various types of DNA, including circular templates, were known. A CPLM is circular and contains

> hairpins and single- and double-stranded regions. Pacific Biosciences' '146 patent disclosed methods of "Single Molecule Sequencing" and expressly stated that its "sequencing methods can be used" with "any nucleic acid molecule, including double-stranded or single-stranded, linear or circular nucleic acids (e.g., circular DNA), single stranded DNA hairpins," and various other structures including "complex nucleic acid structures" such as "tandem repeats."
>
> …
>
> Thus, for a CPLM made by attaching hairpin linker oligonucleotides to target DNA of the length produced by nebulization (700-1330 bp), one skilled in the art would predictably have been able to obtain forward and reverse strand sequences from a single molecule sequencing reaction.

Ex. 2034, ¶ 53. (citations omitted).

In addition, ITRI adduces documentary evidence of the state of the art at the time of the '313 application filing. ITRI's Exhibit 2044[14] is an article entitled *The 3′ to 5′ exonuclease activity of Mre11 facilitates repair of DNA double-strand breaks* by Tanya T. Pauli and Martin Gellert and published in the journal MOLECULAR CELL, Vol. 1, 969–979, in June 1998. The article states, *inter alia*, that: "[T]he dumbbell substrate was composed of two [DNA] hairpins ligated together." Ex. 2044 at 978.

Similarly, ITRI's Exhibit 2037[15] is a European Patent Application, EP 0 653 489 A1, entitled *Coccidiosis vaccin* [sic], dated September 9, 1994. The application discloses: "The remaining cDNA was end-repaired using T4 DNA polymerase in the presence of all four dNTPs at 37°C for 30 minutes. EcoRI adaptors were ligated onto the blunted ends of the cDNA using T4 DNA ligase at 8°C for 24 hours" as an example of blunt-end ligation techniques known in the art. Ex. 2037, p. 11, ll. 24-26.

[14] Paper No. 121
[15] Paper No. 114

ITRI also points to its Exhibit 2038[16], entitled *Accurate whole human genome sequencing using reversible terminator chemistry*, by David R. Bentley et al., which was published in NATURE, vol. 456, pp. 53-59 on November 6, 2008. On pages 1-2 of the supplemental information to the article (published by NATURE), the article teaches fragmentation of purified DNA by nebulization and subsequent ligation using a single-A overhang. Ex. 2038 at 1-2.

Similarly, with respect "obtaining sequence data …by single molecule sequencing," PacBio adduces no substantial evidence in support of its argument that a person of ordinary skill in the art would not have known that ITRI was in possession of the invention disclosed in the '313 application. As noted *supra*, the Zon Declarations are silent with respect to the subject.

In contrast, ITRI points to PacBio's '146 patent, published November 7, 2007, entitled *Apparatus and Method for Analysis of Molecules*, which discloses "single molecule sequencing," which:

> [C]an be used to determine the nucleic acid of any nucleic acid molecule, including double-stranded or single-stranded, linear or circular nucleic acids (e.g., circular DNA), single stranded DNA hairpins, DNA/RNA hybrids, RNA with a recognition site for binding of the polymerase, or RNA hairpins. The methods of the present invention are suitable for sequencing complex nucleic acid structures, such as 5' or 3' non-translation sequences, tandem repeats, exons or introns, chromosomal segments, whole chromosomes or genomes.

Ex. 2036, col. 23, ll. 30; 50-59. Dr. Levy declares that, in addition to the '146 patent, multiple publications had discussed single molecule sequencing, including John Eid et al. *Real-time DNA sequencing from single polymerase molecules*, 323 SCIENCE 133-38

[16] Paper No. 115

(January 2, 2009) (Ex. 2039[17]), Pushpendra K. Gupta, *Single-molecule DNA sequencing technologies for future genomics research*, 26(11) TRENDS IN BIOTECHNOLOGY 602-11 (August 21, 2008) (Ex. 2040[18]), Elaine R. Mardis, *Next-Generation DNA Sequencing Methods*, 9 Ann. Rev. Genomics Hum. Genet. 387-402 (2008) (Ex. 2042[19]), and Jay Shendure & Hanlee Ji, *Next-generation DNA sequencing*, 26(10) NATURE BIOTECHNOLOGY 1135-1145 (October 2008) (Ex. 2046[20]), which teaches:

> [A] highly sensitive fluorescence detection system is used to directly interrogate single DNA molecules via sequencing by synthesis. Template libraries, prepared by random fragmentation and poly-A tailing (that is, no PCR amplification), are captured by hybridization to surface-tethered poly-T oligomers to yield a disordered array of primed single-molecule sequencing templates. *At each cycle, DNA polymerase and a single species of fluorescently labeled nucleotide are added, resulting in template-dependent extension of the surface-immobilized primer-template duplexes*.

Ex. 2046 at 1140 (emphasis added); Ex. 2034, ¶ 46.

Thus, on the record before us, we determine that ITRI has not shown that the '313 application lacks description of an embodiment of the Count when the '313 application is considered from the perspective of a person of ordinary skill in the art.

B. Lack of Enablement

PacBio next argues that the '313 application fails to include a working example pertaining to the subject matter of the count, nor does it include any working examples related to the individual method steps set forth in the count. Motion at 10. According to

[17] Paper No. 116
[18] Paper No. 117
[19] Paper No. 119
[20] Paper No. 123

PacBio, there are, for instance, no working examples pertaining to construction of a CPLM; specifically locking the forward and reverse strands of a nucleic acid sample together to form a circular pair-locked molecule. *Id.* Nor, asserts PacBio, does the '313 application include a working example directed to obtaining sequence data of the CPLM by single molecule sequencing. Motion at 11.

PacBio asserts that, where the claimed invention is the application of an unpredictable technology in the early stages of development, an enabling description must provide a specific and useful teaching. Motion at 11 (citing *Genentech, Inc. v. Novo Nordisk, AIS*, 108 F.3d 1361, l367-l368 (Fed. Cir. 1997); *Chiron Corp. v. Genentech, Inc.*, 363 F.3d 1247, 1254 (Fed. Cir. 2004). PacBio alleges that the '313 application fails to provide any such teaching. *Id.*

ITRI responds that the '313 application provided substantial guidance for making a CPLM by locking forward and reverse strands of a double stranded nucleic acid together. First, argues ITRI, it teaches that a "DNA molecule with high molecular weight is cut into smaller fragments (target DNAs)"; more specifically, "[g]enomic DNA are shredded to smaller fragments (target DNAs)" which are double-stranded, i.e., have forward and reverse strands. Opp. at 17 (citing Ex. 2002, p. 5, ¶¶ 1-2, Fig. 2). According to ITRI, one skilled in the art would have understood that a target DNA (a shredded piece of genomic DNA) could be prepared using a well-known fragmentation technique, e.g., nebulization. *Id.* (citing Ex. 2034, ¶¶ 36, 50; e.g., Ex. 2038, at ¶¶ 1-2; 2045, Abstract).

ITRI argues further that the '313 application then teaches that "[e]ach target DNA is connected together into a circular structure through a series of steps …. Each target DNA is built into a pair-locked molecule construct, where both complementary strands of this target DNA are connected." Opp. at 17 (quoting Ex. 2002 at 5, ¶¶ 3-4). ITRI

repeats its argument *supra* that there were at least three known molecular biology techniques for achieving this. *Id.* (citing Ex. 2034, ¶¶ 39-42, 51; Ex. 2044, at 978; Ex. 2037, p. 11, ll. 22-26; Ex. 1009, at 30-31; Ex. 2038, ¶¶ 1-2 of SI; Motion 1, p. 5, ll. 11-12. ITRI contends, therefore, that the level of experimentation needed to perform these procedures to make a CPLM and the unpredictability associated with them would have been minimal. Opp. at 17.

ITRI argues further that, with respect to single-molecule sequencing, PacBio had already enabled that invention in its '146 patent. Opp. at 18. Furthermore, contends ITRI, a CPLM has no features that would require substantial experimentation to adapt the technique in PBC's '146 patent to a CPLM. Opp. at 18-19 (citing Ex. 2034, ¶¶ 53-54). On the contrary, ITRI argues, one skilled in the art would have correctly predicted that the '146 technique would work with a CPLM. *Id.*

The standard for enablement under 35 U.S.C. § 112 (first paragraph) is whether one of ordinary skill in the art could have made and used the invention without undue experimentation. *See In re Wands*, 858 F.2d 731, 737 (Fed. Cir. 1988). In the context of an interference, enablement is required only for one embodiment within the Count. *Hunt*, 523 F.2d at 1389. Furthermore, the question of enablement is whether the disclosure is sufficient to enable those skilled in the art to practice the claimed invention; therefore, the specification need not disclose what is well known in the art. *In re Myers*, 410 F.2d 420, (CCPA 1969).

PacBio has adduced no persuasive evidence or testimony to show that a person of ordinary skill in the art could not have made and used the invention without undue experimentation. As ITRI has demonstrated *supra*, the limitations of the count that PacBio is contesting were available in the prior art prior to the filing of the '313 patent.

For example, Dr. Levy, ITRI's declarant states that:

> One of ordinary skill would have understood which procedures would have been suitable given the guidance in the '313 application. Accordingly, there would have been little, if any, need for one of ordinary skill in the art to experiment to lock the forward and reverse strands of a nucleic acid sample together to form a circular pair-locked molecule"

and that: "One skilled in the art would also have known how to use well known, predictable techniques for single molecule sequencing to obtain sequence data for the forward and reverse strands of the CPLM in accordance with the '313 application disclosure with little, if any, need for experimentation." Ex. 2034, ¶¶ 52, 54. Moreover, Dr. Levy supports this conclusion by citing substantial evidence; against which PacBio adduces no persuasive evidence or testimony to rebut. *See, e.g.*, Ex. 2034, ¶¶ 36-52. We therefore conclude that PacBio has not met its burden of showing that an embodiment of the Count is not enabled by the '313 disclosure. PacBio's Motion 1 is consequently DENIED.

III. PacBio Motion 2[21] for judgment of unpatentability under 35 U.S.C. § 103

PacBio's Motion 2 seeks judgment against ITRI on the ground that all of the involved claims 1-28 of ITRI's US 8,486,630 (the "'630 patent")[22] would have been obvious in view of prior art under 35 U.S.C. § 103(a). Motion at 1. Specifically, PacBio alleges that the involved claims would have been obvious over the combination of

[21] Paper No. 56
[22] Paper No. 2001

PacBio's US 8,153,375[23] (filed March 27, 2009) (Ex. 1011[24]) in combination with C.D. Laird et al., *Hairpin-bisulfite PCR: Assessing epigenetic methylation patterns on complementary strands of individual DNA molecules*, 101(1) PROC. NAT. ACAD. SCI. 204. (January 6, 2004) (Ex. 1012[25]) ("Laird") or T. Matsumura et al., *Photochemical transition of 5-methylcytosine to thymine by DNA photoligation*, NUCLEIC ACIDS SYMPOSIUM SERIES NO. 51, 233 (Oxford Univ. Press, 2007) (Ex. 1013[26]) ("Matsumura").[27]

PacBio argues that independent claims 1, 24, and 26 are all based on the use of a CPLM to obtain the sequence of the forward and reverse strands of a double stranded nucleic acid, and thereafter to determine the positions of modified bases in the double-stranded nucleic acid. Motion at 4 (citing Ex. 1002). Claim 1 of the '630 patent is exemplary:

---

[23] (The '375 patent was filed March 27, 2009 and issued April 10, 2012). The application which issued as the '375 patent, US 2009/0298075 Al, was published on Dec. 3, 2009. The '375 patent claims priority to provisional applications 61/099,696, filed September 24, 2008 (Ex. 1006), and 61/072,160, filed March 28, 2008 (Ex. 1007) (respectively, "the '696 application" and "the '160 application").

[24] Paper No. 46

[25] Paper No. 47

[26] Paper No. 48

[27] PacBio also argues that, if PacBio's Motion 1 were granted, then ITRI should receive only the benefit of the filing date of the '630 patent, which is November 5, 2009. Motion at 3. PacBio argues that, if it should prevail on Motion 1, PacBio's Cold Spring Harbor Personal Genomes Meeting Presentation (October 12, 2008) (Ex. 1009) (Paper No. 44) ("the Personal Genomes presentation") would qualify as a prior art reference under 35 U.S.C. § 102(b). *Id.* As we have related *supra*, PacBio's Motion 1 is denied because we determined that PacBio did not show that the ITRI benefit application lacked a constructive reduction to practice of the Count. Under the circumstances before us, and as PacBio has not advanced an argument that the Personal Genomes Meeting Presentation is prior art if Motion 1 is denied, we have not considered it further in our decision.

Interference 105,970

1. A method of determining a sequence of a double-stranded nucleic acid sample and a position of at least one modified base in the sequence, comprising:

a. locking the forward and reverse strands together to form a circular pair-locked molecule;

b. obtaining sequence data of the circular pair-locked molecule via single molecule sequencing, wherein the sequence data comprises sequences of the forward and reverse strands of the circular pair-locked molecule;

c. determining the sequence of the double-stranded nucleic acid sample by comparing the sequences of the forward and reverse strands of the circular pair-locked molecule;

d. altering the base-pairing specificity of bases of a specific type in the circular pair-locked molecule to produce an altered circular pair-locked molecule;

e. obtaining the sequence data of the altered circular pair-locked molecule wherein the sequence data comprises sequences of the altered forward and reverse strands; and

f. determining the positions of modified bases in the sequence of the double-stranded nucleic acid sample by comparing the sequences of the altered forward and reverse strands.

Ex. 2001. The ’630 patent teaches a CPLM constructed by ligating hairpin inserts to both ends of a double-stranded nucleic acid. Motion at 4.

PacBio asserts that, more than a year prior to the benefit currently accorded to ITRI, PacBio filed the ’160 application, which teaches ligating hairpin adaptors to the ends of a double-stranded DNA fragment to generate a CPLM (the “SMRTbell”™ template”), allegedly intended for use as a template for single molecule real-time

("SMRT") sequencing. Motion at 4-5. According to PacBio, a SMRTbell™ has the same structure and components as a CPLM in that it contains a double-stranded region flanked by two hairpins and, further, that polymerase extension from a primed SMRTbell™ template, as for a CPLM, generates sequence information for both the forward and reverse strands. Motion at 5.

PacBio contends that although the '375 patent discloses the use of a CPLM for sequencing the forward and reverse strands of a single molecule double-stranded nucleic acid template, neither explicitly discloses the use of the template to determine the position of a modified base, or altering the base pairing specificity of a specific type of base in the template and thereafter determining the position of the modified base, as recited in ITRI's claims 1, 24 and 26. Motion at 5. However, argues PacBio, ITRI's '630 patent acknowledges that such methods were known in the art. *Id.* (citing Ex. 1002, cols. 22-24). Specifically, the '630 patent describes Laird's method of converting cytosine (but not 5-methylcytosine) to uracil via bisulfite treatment and Matsumura's photochemical conversion of 5-methylcytosine (but not unmodified cytosine) to thymine. *Id.* at 6 (citing Ex. 2002, col. 23, ll. 57-65; col. 23, ll. 49-56). PacBio asserts that both base conversion methods result in a modified base that is paired with a base having a base-pairing specificity different from the preferred partner base of the modified base.[28]

PacBio argues that it would have been obvious at the time that ITRI's '313 application was filed (April 7, 2009) to construct and determine the sequence of a double-stranded nucleic acid in a CPLM as taught by the '375 patent and to subject the CPLM to either (i) bisulfite conversion or (ii) photochemical conversion to create a

[28] For example, a cytosine converted to uracil would be found to be base-paired with guanine rather than the expected adenine.

modified base by which the position of a 5-methylcytosine residue could be determined by comparing the sequences of the forward and reverse strands, as recited in claims 1, 5, 6, and 24-26 of ITRI's '630 patent. Motion at 6.

According to PacBio, one of ordinary skill in the art would have been motivated to compare the sequences of the forward and reverse strands because the '375 patent explicitly teaches that comparing the sequence of the forward and reverse strand is one of the advantages of using a CPLM, e.g., to provide redundant sequence information. *Id.* PacBio argues that one of ordinary skill would have been further motivated to compare the forward and reverse strands in order to increase the accuracy of 5-methylcytosine determination, given the well-known importance of DNA methylation in the regulation of gene expression, genomic imprinting and X-chromosome inactivation. *Id.* at 6-7 (citing Ex. 1002, col. 2, ll.14-23).

Finally, argues PacBio, an artisan of ordinary skill would have had a reasonable expectation of success in determining the location of a modified base in a CPLM, given that the techniques for determining the location of 5-methylcytosine residues were known, and the claimed methods simply involve adding or employing the unique CPLM template in this known sequencing reaction. *Id.* at 7 (citing Ex. 1002, col. 23, ll.49-65, Abstract; Ex. 1013, Abstract). PacBio contends that because the CPLM sequencing template, single molecule sequencing and base conversion of a sequencing template had previously been disclosed in the prior art, ITRI's invention amounts to a simple combination of prior art elements according to known methods to achieve predictable results.

PacBio further contends that claims 2-4 are and 7-23 are all dependent on claim 1, and incorporate all of its limitations. Motion at 8. Moreover, argues PacBio, the

additional limitations recited in the dependent claims are also found in the '375 patent and therefore do not provide a basis for patentability.[29]

PacBio also argues that claims 27-28 are also based on the use of a CPLM to obtain the sequence of the forward and reverse strands of a double-stranded nucleic acid, and thereafter to determine the positions of modified bases in the double-stranded nucleic acid as discussed for the other claims above. Motion at 13. Claim 27 is exemplary and recites:

> 27. A method of determining a sequence of a double-stranded nucleic acid sample and a position of at least one modified base in the sequence, comprising:
>
> a. locking the forward and reverse strands together to form a circular pair-locked molecule;
>
> b. obtaining sequence data of the circular pair-locked molecule via single molecule sequencing, wherein the sequence data comprises sequences of the forward and reverse strands of the circular pair-locked molecule;
>
> c. determining the sequence of the double-stranded nucleic acid sample by comparing the sequences of the forward and reverse strands of the circular pair-locked molecule·
>
> d. obtaining sequencing data of the circular pair-locked molecule via single molecule sequencing, wherein at least one nucleotide analog that discriminates between a base and its modified form is used to obtain sequence data comprising at least one position wherein the at least one differentially labeled nucleotide analog was incorporated; and

[29] PacBio argues separately with respect to claims 2, 4, 7-10, 13, 14, 17-23, and 27-28. Motion at 8-13. ITRI does not respond to these individual arguments in its opposition.

> e. determining the positions of modified bases in the sequence of the double-stranded nucleic acid sample by comparing the sequences of the forward and reverse strands.

Ex. 2001. PacBio points out that, rather than employ a difference in base pairing specificity (as in claims 1-26), claims 27 and 28 refer to the use of nucleotide analogs that discriminate between a base and its modified form. Motion at 13.

According to PacBio, the only difference between these two independent claims is the number of steps in the sequence determination, i.e., claim 27 employs a preliminary sequencing step prior to sequencing using discriminating bases, whereas claims 28 does not. *Id.* Furthermore, argues PacBio, although ITRI was accorded benefit of the '313 application, the subject matter for claims 27 and 28 was not included in the application, and was added only in ITRI's U.S. Application No. 12/613,291 (the application that matured into the '630 patent). *Id.* (citing Exs. 1002; 1003).

PacBio asserts that the '375 patent teaches the use of a CPLM for single molecule sequencing, which ITRI admits in its '313 priority application is the key point of novelty. Motion at 13-14 (citing Ex. 1003, p. 3). Otherwise, contends PacBio the sequencing techniques disclosed and claimed in the '630 patent were previously known to one of ordinary skill in the art: in particular, the use of discriminating nucleotide analogs to identify modified bases, including 5-methylcytosine bases, was disclosed in US 7,399,614 patent (the "'614 patent"), as acknowledged in the '630 patent. *Id.* at 14 (citing Ex. 1002, col. 24, ll. 42-46; Ex. 1014[30], Abstract, claims 1-29).

[30] Paper No. 49

According to PacBio, it would have been obvious at the time that the '630 patent was filed to construct and determine the sequence of a double-stranded nucleic acid in a CPLM, as disclosed in the '375 patent, and to include discriminating nucleotide analogs in the sequencing reaction as disclosed by the '614 patent, to identify the position of modified bases as recited in claims 27-28. Motion at 14. PacBio asserts that one of ordinary skill would have been motivated to do so to increase the accuracy of the sequence determination, given the importance of DNA methylation in the regulation of gene expression, genomic imprinting, and X-chromosome inactivation. *Id.* Further, PacBio contends that the ordinarily skilled artisan would have had a reasonable expectation of success, given that discriminating analogs for determining the location of 5-methylcytosine residues were known, and the claimed methods simply involve employing the unique CPLM template. *Id.*

ITRI responds that independent claims 1, 24, and 26 require using a disagreement between the forward and reverse strand sequences of the CPLM to determine a modified base's position. Opp. at 2. According to ITRI, claim 24, which it chooses as an exemplar, recites determining the position of at least one modified base in the sequence of a nucleic acid sample "by comparing the sequences of the forward and reverse strands" and that at least one modified base in the double stranded nucleic acid sample "is paired with a base having a base pairing specificity different from its preferred partner base," which is referred to as being "mismatched." *Id.* at 2-3. ITRI argues that claim 24 thus teaches sequencing the forward and reverse strands of a CPLM containing a mismatched modified base, such as a uracil paired to guanine instead of adenine, and that obtaining these sequences reveals a disagreement which is used to conclude that a modified base

was present at the position of the disagreement. *Id.* at 3 (citing, e.g., Ex. 2048[31], ¶¶ 31-33).

ITRI contends that it would be unreasonable to interpret claim 24 as covering a determination of the position of the modified base that does not use the disagreement between the two strands of DNA, because the claim requires a mismatched modified base in the double stranded nucleic acid sample. Opp. at 4 (citing Ex. 2048, ¶ 34). According to ITRI's declarant, Dr. Levy:

> One of ordinary skill would not understand claim 24 as meaning that the disagreement between the sequence data of the forward and reverse strands can be disregarded or that a modified base position can be determined using only some other, undefined information independent of the disagreement. Interpreting claim 24 to mean that a modified base position can be determined using only some other, undefined information is unreasonable because it would render the mismatch irrelevant and superfluous to the determining step in which it appears.

Ex. 2048, ¶ 34. ITRI thus argues that it would not be reasonable to disregard the disagreement, treat the mismatch as superfluous and irrelevant to the step that recites it, and use a different basis—obtained by comparing the strands in some way not defined by claim 24—to deduce a modified base position. Opp. at 4.

ITRI contends that PacBio provides no evidence supporting the attorney argument that its proposed combination of documents would have rendered ITRI's claims obvious to one of ordinary skill in the art. Opp. at 7. ITRI points out that the Zon Declaration takes no position concerning the alleged obviousness of claims 1, 24, and 26. *Id.* (citing Ex. 1010, ¶¶ 9-18).

[31] Paper No. 134

ITRI also argues that PacBio's arguments fail to combine the teachings of its cited prior art references in a manner that shows all of the claimed steps, including determining the position of the modified base using a disagreement between the forward and reverse strand sequences of the CPLM. Opp. at 7. ITRI points out that PacBio has failed to submit a claim chart comparing the claim language to the prior art in conformance with Bd.R. 41.121(e), as well as to construe the claims. ITRI argues that PacBio also fails to explain why one of ordinary skill in the art would have modified the methods of its '375 patent in a manner inconsistent with the teachings of those references.

ITRI agrees with PacBio's contention that the '375 patent teaches construction of a CPLM and use single molecule sequencing to provide redundant sequence information by comparing the forward and reverse strand sequences of the CPLM. Opp. at 7 (citing Motion at 6, ll. 19-21, 26-29). ITRI admits that the '375 patent described "consensus sequence determination through the sequencing of both the sense and antisense strand" of a double stranded segment in a circular molecule. *Id.* (citing Ex. 1011 at 9, ll. 66-67). ITRI contends that, in 2009, the term "consensus sequence" meant a sequence determined from agreement between related positions in a set of sequence data and that confirming sequence data from one strand using its consistency (i.e., its proper Crick-Watson matching) was conventional in the field. Opp. at 8 (citing Ex. 2048, ¶¶ 48, 54; Ex. 2013, 1:21-27 ("Forward and reverse strand sequencing provides the researcher with more information and allows the researcher to evaluate the quality and reliability of the data from both strands. If the bases on both strands complement each other as expected, then this helps to confirm the reliability of the sequence information"). Consequently, argues ITRI, one of ordinary skill would have known that obtaining a consensus sequence with forward and reverse strand sequence data involves looking for complementarity, i.e.,

agreement or redundancy, and discarding disagreements. *Id.* (citing Ex. 2048, ¶¶ 48-49, 54).

Therefore, argues ITRI, the '375 patent's teachings to increase sequencing accuracy by disregarding disagreements between the two strands would have led one skilled in the art away from using disagreements between forward and reverse strands as set forth in the methods of ITRI's claims 1, 24, and 26. Opp. at 9. (citing Ex. 2048; ¶¶ 55, 57, 63). According to ITRI, a method that disregards any mismatch, such as the one disclosed in the '375 patent, is the antithesis of the method recited in claims 1, 24, and 26 of ITRI's '630 patent. *Id.*

ITRI additionally argues PBC failed to explain why Laird or Matsumura would have countered that teaching to arrive at the claimed methods, and that one of ordinary skill in the art would have had no reason to modify the teaching of the '375 patent to reach the methods of any of claims 1, 24, or 26. ITRI contends that neither Laird nor Matsumura teaches or suggests determining the position of a modified base by comparing sequences of forward and reverse strands or by using a disagreement between those strands. Opp. at 10 (citing Ex. 2048, ¶¶ 59-62).

With respect to claim 23, ITRI argues that PacBio's argument fails because PacBio has failed to allege that step (h) was disclosed in any of the references. Opp. at 11. ITRI argues that step (h) recites, in part, "accepting or rejecting at least four of the repeats of the sequence of the nucleic acid sample contained in the sequence data according to the scores of one or both of the sequences of the inserts immediately upstream and downstream of the sample sequences …." *Id.* at 11-12. According to ITRI, PacBio never mentions accepting or rejecting repeats of a sample according to the score of an insert in its argument about claim 23. *Id.* at 12 (citing Motion at 12-13, ll. 24-10). Instead, argues

ITRI, PacBio refers to "scoring methods whereby calls of iterative reads or from complementary sequences are scored by the number of occurrences." *Id.* (citing Motion at 13, ll. 8-9). ITRI argues that the passages PacBio cites in Exs. 1011 and 1006 also do not disclose accepting or rejecting repeats of a sample according to the score of an insert. *Id.*

With respect to claims 27 and 28, ITRI argues that the '375 patent neither teaches nor suggests using a nucleotide analog that discriminates between a base and its modified form or determining the positions of modified bases. Opp. at 13 (citing Ex. 2048, ¶ 66). According to ITRI, PacBio provides only an allegedly conclusory allegation that a person "of ordinary skill would have been motivated to do so to increase the accuracy of the sequence determination, given the importance of DNA methylation in the regulation of gene expression, genomic imprinting and X-chromosome inactivation." *Id.* (citing Motion at 14, ll. 11-14). ITRI argues that the '375 patent provide increased accuracy, and PacBio has not shown or argued that its alleged combination would provide any further increase. *Id.* (citing Motion at 14, ll. 7-18; Ex. 2048, ¶ 56). Furthermore, argues ITRI, even if there were a marginal increase in accuracy under some circumstances, PacBio has not shown that this motivation would outweigh the expected problem of polymerase enzyme inhibition. *Id.* (citing Motion at 13-14, ll. 15-18).

ITRI argues that PacBio's alleged motivation requires that one of ordinary skill must have understood that combining the '614 patent with the '375 patent would have "increase[d] the accuracy of the sequence determination" beyond what the '375 patent already provided. Opp. at 14. However, argues ITRI, PacBio's statement alleging this motivation lacks support from any evidence or even further argument. *Id.*

We are persuaded by PacBio's arguments and conclude that the claims of the '630 patent are obvious over the combination of the '375 patent, Laird and Matsumara. Both parties agree that the '375 patent teaches construction of a CPLM and the use of single molecule sequencing to provide redundant sequence information by comparing the forward and reverse strand sequences of the CPLM. *See* Opp. at 7; Motion at 6. Furthermore both parties agree that the '375 patent described "consensus sequence determination through the sequencing of both the sense and antisense strand" of a double stranded segment in a circular molecule. *Id*. Consequently, this leaves remaining only limitation c of claim 24 of the '630 patent, which ITRI has chosen as an exemplar. Limitation c of claim 24 recites:

> c. determining the sequence of the double stranded nucleic acid sample and the position of the at least one modified base in the sequence of the double stranded nucleic acid sample by comparing the sequences of the forward and reverse strands of the circular pair-locked molecule wherein at least one modified base in the double-stranded nucleic sample is paired with a base having a base pairing specificity different from its preferred partner base.

Ex. 1002. ITRI's argument that one of ordinary skill would have known that obtaining a consensus sequence with forward and reverse strand sequence data involves looking for complementarity, i.e., agreement or redundancy, and discarding disagreements is countered by the teachings of Laird.

Laird teaches that methylated and unmethylated CPG dyads in a bisulfite-treated DNA sequence can be identified by the matching or mismatching of cytosines in the forward and reverse strand sequence data. Laird recites:

> Bisulfite conversion was used because it provides information on the methylation state of individual cytosines by converting cytosine (but not 5-

> methylcytosine) to uracil, and subsequently to thymine upon PCR amplification (28, 29). Bisulfite conversion also reduces base pair complementarity within the hairpin, a reduction that we found useful for PCR amplification, as well as for high-fidelity cloning and sequencing of hairpins containing unmethylated cytosines.

Ex. 1012 at 2. Moreover, the purpose of Laird's study is to examine the stability of cytosine methylation in CpG/CpG[32] dyads, which it achieves by comparing the forward and reverse strands of the DNA sample and identifying "mismatched" base pair configurations. Figure 2D of Laird is illustrative:

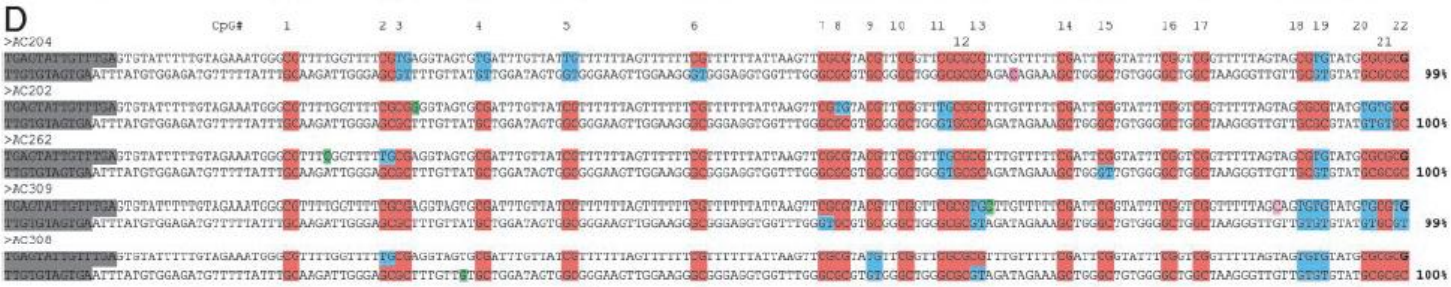


Figure 2D depicts forward and reverse strands of bisulfite-PCR treated DNA

The highlighted sequences of the paired forward and reverse strands depicted in Figure 2D of Laird depict CpG sequences that have either been transformed to blue-highlighted thymine-guanine pairs (indicating non-methylated cytosine was originally at that locus) and red-highlighted cytosine-guanine pairs (indicating that the unconverted cytosine is 5-methylcytosine at that locus). Although the principle aim of Laird's teaching is to investigate the stability of methylated CpG/CpG dyads, the manner in which this aim is achieved is by comparing the sequences of the converted forward and

[32] CpG refers to covalently linked (i.e. positioned adjacently in a strand) cytosine and guanine bases.

reverse DNA strands and identifying mismatched (i.e., thymine-guanine) base pairs in the sequence. *See* Ex. 1012 at 3 ("In sequences with efficient conversion of non-CpG cytosines (97% conversion), the four informative hemimethylated CpG/CpG dyads (i.e., those between the primer sites) showed no conversion of the methylated CpGs (0/24), and 100% conversion (24/24) of the four complementary, unmethylated CpGs").

ITRI argues that the '375 patent would teach a person of ordinary skill in the art to look for agreement in the sequencing of base pairs in the forward and reverse strands and to discard disagreements (i.e., mismatched base pairs). ITRI implicitly argues, therefore, that the '375 patent teaches away from the methods taught by the '630 patent.

We disagree. A reference teaches away when "a person of ordinary skill, upon reading the reference, would be discouraged from following the path set out in the reference, or would be led in a direction divergent from the path that was taken by the applicant." *In re Gurley*, 27 F.3d 551, 553 (Fed. Cir. 1994). ITRI points to no teaching in the '375 patent, or the other cited prior art references, that would explicitly or implicitly discourage a person of ordinary skill from looking for base pair disagreements as well as base pair agreements and, as we have related *supra*, Laird explicitly teaches looking for guanine-thymine mismatches as evidence of non-methylated cytosine in a forward or reverse strand locus. Moreover, observing disagreement between base pairs is inherent in observing base pair agreement, i.e., base pair agreement/disagreement under the Crick-Watson rules is a binary choice; the bases either or properly matched or they are not. Thus, when observing whether the base pairs agree, one necessarily observes whether they disagree. *See In re Cruciferous Sprout Litig.*, 301 F.3d 1343, 1349 (Fed. Cir. 2002) (Inherency can only be established when "prior art necessarily functions in accordance with, or includes, the claimed limitations").

With respect to claim 23, we are persuaded that the cited prior art references teach or suggest limitation h, as ITRI maintains. Limitation h recites:

> h. accepting or rejecting at least four of the repeats of the sequence of the nucleic acid sample contained in the sequence data according to the scores of one or both of the sequences of the inserts immediately upstream and downstream of the sample sequences, subject to the condition that at least one sample sequence in each orientation is accepted.

Ex. 1002. The '375 patent recites:

> Although referred to herein as comparing or assembling the sequence data from multiple reads of a given sequence, and/or from the sense and antisense strands of the sequence, it will be appreciated that any method of assigning a consensus determination to a particular base call from multiple reads of that position of sequence, and/or to a provide an overall consensus sequence for that segment, will be envisioned and encompassed by the term "compare."

Ex. 1011, col. 11, ll. 40-47. The '375 patent thus explicitly teaches determining consensus from "multiple reads of that position of sequence." Moreover, Laird teaches determining sequences and agreement/disagreement of base pairs at given loci. *See* Zon Decl. 2 at Fig. 2. Although neither reference explicitly requires that at "least four of the repeats" are accepted or rejected, the very nature of the teachings of the '375 patent and Laird will produce at least four repeats. The '630 patent defines "repeat" as:

> A repeat or repeat sequence is a sequence that occurs more than once in a nucleic acid. When repeats are present in a nucleic acid molecule, all instances of the sequence, including the first instance, are considered repeats. Repeats include sequences that are reverse complements of each other, such as occur in a circular pair-locked molecule.

Ex. 1002, col. 10, ll. 40-44. Thus a single iteration of the formation of the CPLM and single replication of the double stranded DNA sample will yield four repeats:

the forward and reverse strands of the original sample and the complementary strands formed by polymerization. *See, e.g.*, Ex. 1011, col. 20, ll. 50-62, Figs. 5, 6. We therefore conclude that the combined cited prior art references teach limitation h of claim 23.

With respect to claims 27 and 28, we are similarly not persuaded by ITRI's argument that the prior art fails to teach or suggest the limitation reciting "[a]t least one nucleotide analog that discriminates between a base and its modified form is used to obtain sequence data comprising at least one position wherein the at least one differentially labeled nucleotide analog was incorporated." ITRI argues that the '375 patent provides increased accuracy, and asserts that PacBio has not shown or argued that its alleged combination would provide any further increase. Opp. at 13.

Laird teaches such a method or discriminating between methylated and non-methylated cytosines, and comments extensively on the developmental importance of cytosine methylation, particularly with respect to epigenetic cellular inheritance. *See, e.g.*, Ex 1025 at 1. We find that a person of ordinary skill in the art would know that the teachings of the '375 patent could be combined with the teachings of Laird to increase the accuracy of the method by discriminating between methylated and non-methylated cytosines, and we conclude that the combined cited prior art references tech or suggest all of the limitations of claims 27 and 28.

"The combination of familiar elements according to known methods is likely to be obvious when it does no more than yield predictable results." *KSR Int'l Co. v. Teleflex Inc.*, 550 U.S. 398, 416 (2007). The combined cited prior art references teach or suggest all of the limitations of the claims of the '630 patent and we

conclude that a person of ordinary skill in the art would be motivated to combine the references to increase the accuracy of the invention, particularly with respect to discriminating between methylated and demethylated cytosine bases. PacBio's Motion 3 for judgment of unpatentability under 35 U.S.C. § 103(a) against ITRI's '630 patent is consequently GRANTED.

IV. ITRI'S Motion 2 to deny PacBio the benefit of its Application No. 61/201,551[33]

ITRI's Motion 2 seeks judgment that Pacific Biosciences is not entitled to the benefit of its Application No. 61/201,551 (the "'551 application")[34] (Ex. 2026) which was filed December 11, 2008. Motion[35] at 1. PacBio's involved US Applications 13/633,673 and 13/930,178 claim the benefit of the '551 application. Redecl.[36] at 3. According to ITRI, the '551 application fails to provide any disclosure indicating that PacBio invented an embodiment within the scope of the Count. *Id.*

Specifically, ITRI alleges that, in contrast to the method disclosed in ITRI's '630 patent, claim 24 of which constitutes the sole count of this interference, the '551 application does not use a mismatch or a disagreement between bases pairs in the forward and reverse strands. Motion at 5 (citing Levy First Decl., ¶ 36) ("Levy Decl. 1").

[33] Per the March 31, 2014 order of the Board, ITRI's Motion 3 is deferred until the priority phase of the interference. *See* Paper No. 106.
[34] Paper No. 84
[35] Paper No. 59
[36] Paper No. 28

Rather, contends ITRI, the '551 application describes an approach using a distinct signal that occurs when a nucleotide is incorporated across from a modified base during sequencing-by-synthesis. *Id.* (citing Ex. 2026, ¶¶ 16-17; Ex. 2003, ¶ 33).

ITRI argues that, when interpreted in light of ITRI's '630 Specification, the Count requires using a disagreement between forward and reverse strand sequences of the CPLM to determine the position of a modified base. Motion at 6 (citing Levy Decl. 1, ¶ 30). The count of the instant interference recites, in relevant part:

> c. determining the sequence of the double stranded nucleic acid sample and the position of the at least one modified base in the sequence of the double stranded nucleic acid sample by comparing the sequences of the forward and reverse strands of the circular pair-locked molecule wherein at least one modified base in the double-stranded nucleic sample is paired with a base having a base pairing specificity different from its preferred partner base [i.e., is "mismatched"].

Ex. 1002. Therefore, argues ITRI, the count comprises sequencing the forward and reverse strands of a CPLM containing a mismatched modified base, such as a uracil paired to guanine, and that obtaining these sequences reveals a disagreement—the uracil directs incorporation of an adenine into the growing strand, whereas the guanine directs incorporation of a cytosine into the opposite growing strand. Motion at 7 (citing Levy Decl. 1, ¶¶ 28-29). In the limitation of the count recited *supra*, the disagreement is used to conclude that a modified base was present at the position of the disagreement (specifically, that the modified base was a cytosine that has been deaminated to uracil). *Id.* (citing Levy Decl. 1, ¶ 30).

ITRI argues that the '551 application does not describe an embodiment of this limitation of the count. Motion at 8. Specifically, argues ITRI, the '551 application lacks

a written description of "determining … the position of the at least one modified base … by comparing the sequences of the forward and reverse strands of the circular pair-locked molecule, wherein at least one modified base in the double stranded nucleic acid sample is" mismatched. *Id.* (citing Levy Decl. 1., ¶¶ 37, 39). ITRI contends that the '551 Specification lacks express description of comparing forward and reverse strand sequences that disagree at a modified base position due to a mismatch involving the modified base, and using that mismatch to determine the modified base position. *Id.* at 9 (citing Levy Decl. 1, ¶¶ 36-37).

According to ITRI, the only mention of using the reverse strand sequence in a modified base determination in the '551 application recites "For example, sequence reads from the sense or 'forward' strand can be compared to sequence reads from the antisense or 'reverse' strand for the same nucleic acid template to further validate the existence of one or more modified bases in the template nucleic acid." *Id.* at 9 (quoting Ex. 2026, ¶ 17; citing Levy Decl. 1, ¶ 36). ITRI contends that neither that sentence, nor its context, conveys using a disagreement between the forward and reverse strand sequences to determine the modified base position and, therefore, the '551 specification does not expressly describe using such a disagreement to determine the position of the modified base. *Id.* (citing Levy Decl. 1, ¶ 36).

Similarly, argues ITRI, neither paragraph 18 nor 21 of the '551 application supports the language of the count reciting a mismatch in the sample. Motion at 9. (citing Ex. 2010 at 5-6). According to ITRI, Paragraph 18 refers to several modifications "to improve the signal discrimination approaches provided herein," including the use of modified nucleotides to create the new strand during sequencing "to enhance the discrimination of various bases" but does not disclose methods of discriminating

modified bases that involve a mismatch in the sample. *Id.* (citing Ex. 2026, ¶ 18). ITRI observes that paragraph 21 recites: "[a]ltered nitrogenous bases may exhibit different base pairing and/or base stacking properties as compared to native bases. For example, 'universal bases' can pair with all four native bases." *Id.* at 10 (quoting Ex. 2026, ¶ 21). ITRI contends that different base pairing properties, particularly those of universal bases, is not the same thing as a mismatch. *Id.* (citing Levy Decl. 1, ¶ 70).

Further, ITRI notes that PacBio's Application No. 61/099,696[37] (the "'696 application," filed September 24, 2008) (Ex. 2015), which is incorporated by reference into the '551 application, recites "comparing sequence data from each strand to determine a consensus sequence for the double stranded segment" where the double stranded segment was sequenced as part of a circular molecule. Motion at 11 (quoting Ex. 2015, p. 30). ITRI also observes that the '696 application recites "as with circular templates, the template configurations of the invention provide single molecular consensus sequences, where sequencing a given template provides duplicative data of the sequence information obtained, and thereby improves accuracy over linear templates by providing confirming or consensus sequence data on a given sequence." *Id.* (quoting Ex. 2015, ¶ 33). However, argues ITRI, the consensus sequence is obtained by obtaining duplicative or redundant data that shows agreement between the forward and reverse strands, and any mismatches at a given position in the sequencing data will be overruled by the more prevalent redundant agreement shown in the data. *Id.* (citing Levy Decl. 1, ¶¶ 47-48).

ITRI also argues that the '551 application's use of the term "validation" demonstrates that the '551 inventors did not contemplate that the forward and reverse

[37] Paper No. 75

strand sequences would show a disagreement. Motion at 12 (citing Ex. 2026). "Validation" of a modified base with reverse strand sequence data, contends ITRI, implies that the data show complementarity, i.e., agreement or redundancy, with the modified base. *Id.* (citing Levy Decl. 1, ¶ 61). ITRI argues that as of December 11, 2008, sequencing approaches for identifying modified bases using a sequence comparison would compare one strand of a sample sequence to a reference sequence, not the reverse strand of the nucleic acid. *Id.* (citing Ex. 2020, p. 206; Ex. 2003, ¶ 60).

Finally, ITRI argues that whereas the count requires a mismatched modified base in the sample and determining the position of the modified base using a disagreement between forward and reverse sequences, statements about its inventive detection of modified bases, particularly uracil, in the '551 application explicitly exclude methods of determining the position of a modified base using a disagreement between the forward and reverse strand sequences at the modified base position. Motion at 14 (citing Levy Decl. 1, ¶ 74).

ITRI points to the '551 applications recitation that uracil "can be found in DNA … as the result of bisulfite-conversion of cytosine in a common protocol used to discriminate methylated cytosine through DNA sequencing. Unlike this common protocol, the methods herein directly detect the modified base rather than relying on the similarity of uracil to thymine." *Id.* at 14-15 (quoting Ex. 2026, ¶ 23). ITRI contends that the methods of the '551 application rely on detecting some behavior of uracil that differs from that of thymine, such as a change in the kinetics of sequencing-by-synthesis data, a difference in behavior of a polymerase when it encounters uracil, or statistical

differences in error rate frequencies, none of which uses a mismatch to detect uracil. *Id.* at 15 (citing Levy Decl. 1, ¶¶ 73-74, Ex. 2026, ¶¶ 16-17, 28, 29).

PacBio responds that ITR's arguments are based on an overly restrictive and incorrect interpretation of the count and that a step of "using a disagreement" is not recited in the count. Opp. at 4. According to PacBio, nowhere in the count is it recited that the method requires using a disagreement to determine the position of a modified base, let alone using the disagreement at all, or does the count recite that "the at least one modified base" detected by comparing the forward and reverse strands is the same modified base as that recited in the last limitation of the count. *Id.* at 5.

PacBio contends that, upon considering the '630 Specification in its entirety, one of ordinary skill in the art would have understood that the Specification teaches detecting the position of a modified base via the method of the count regardless of whether the base is mismatched. Opp. at 5 (citing Ex. 1002, cols. 34-35, ll. 60-20; Zon Decl. 1[38], ¶¶ 43-49, 52-53, 57, 61 (Ex. 1024); Zon Decl. 2, ¶¶ 23-27). PacBio points out that in the context of a bisulfite treated DNA template, it is the non-mismatched 5-methylcytosine positions that one of ordinary skill in the art generally seeks to detect and that the position of a 5-methylcytosine can be detected in a bisulfite-treated CPLM by comparing the forward and reverse strands of a CPLM and detecting a normal Crick-Watson C-G base pair. *Id.* (citing Ex. 1002, col. 2, ll.14-23, cols. 34-35; ll. 60-20; Zon Decl. 1, ¶ 47; Zon Decl. 2, ¶¶ 18, 24-26). PacBio argues that ITRI's proposed construction of the count would exclude the detection of a modified base that is thus not mismatched. *Id.*

PacBio also disputes ITRI's interpretation of the claim language to mean that determining the position of a modified base does not encompass validation of the position

[38] Paper No. 136

of the modified base. Opp. at 6. ITRI contends that, although "determining" is not defined in the '630 patent, its use in the specification is 27 consistent with the ordinary and customary meaning of the term. *Id.* PacBio points to Example 6 of the '630 Specification, which describes comparing the forward and reverse strands of a CPLM to "determine the sequence of the nucleic acid sample and the position of the at least one 5-methylcytosine" and that the template nucleic acid can also include a mismatched base, for example a uracil mismatched with a guanine, due to bisulfite. *Id.* at 6-7 (quoting Ex. 1002, col. 34-35 ll. 60-20). PacBio contends that this comparing step is a validation of the initial determination of the position of a 5-methylcytosine residue, because the position of the 5-methylcytosine residue is immediately determined upon obtaining a G read in the sequencing data in the absence of a comparing step. *Id.* at 7 (citing Zon Decl. 2, ¶¶ 24, 26). Accordingly, argues PacBio, the act of comparing the two strands to determine the position of a C-G matched pair is a validation of the determination of the position of the 5-methylcytosine, since the position of the 5-methylcytosine is initially determined from reading either the forward or reverse strand alone. *Id.* (citing Zon Decl. 1, ¶ 47; Zon Decl. 2, ¶ 26).

PacBio next argues that CPLM templates and bisulfite treatment are described in the '551 application in the same paragraph that also teaches comparing the forward and reverse strands of the CPLM to determine the presence of a modified base. Opp. at 10. According to PacBio, one of ordinary skill in the art would have readily understood that the sequencing methods taught in the '551 application could be carried out on the bisulfite treated template also described therein and, further, would have also understood that bisulfite treatment could be applied to CPLM templates to create mismatched modified bases (i.e., uracil paired with guanine) and would also leave the base pairing

specificity of 5-methylcytosine (a modified base) unchanged. *Id.* (citing Zon Decl. 1, ¶¶ 40-42).

PacBio also contends that the '551 application further states that a comparison of data reads of the forward and reverse strands "can provide additional information for discriminating modified from unmodified 6 nucleotides." Opp. at 11 (citing Ex. 1004, ¶ [0017]). PacBio contends that one of ordinary skill in the art would have understood that the comparison step taught at ¶ [0017] of the '551 application could be used to determine the position of a mismatched modified base, i.e., uracil, in a bisulfite treated CPLM, initially determined by a separate mechanism such as incorporation signal strength, signal duration or signal frequency. *Id.* (citing Zon Decl. 1 ¶¶ 47, 61).

Accordingly, argues PacBio, when ¶ [0017] of the '551 application is read and interpreted in the context of the '551 application as a whole, one of ordinary skill in the art would understand that it describes an embodiment where comparing the forward and reverse strands of a bisulfite-treated CPLM is used to determine the position of the modified base (e.g., uracil or 5-methylcytosine) determined initially by the other readout mechanisms described in 16 the '551 application, such as signal duration and signal strength. Opp. at 11 (citing 1004, ¶¶ [0016]-[0017], [0026]; Zon Decl. 1, ¶¶ 47, 49, 61).

We are not persuaded by ITRI's arguments. As we related *supra*, a person of ordinary skill in the art as of 2009 would have possessed a Ph.D. in molecular biology, or a related field, and 1-2 years' relevant postdoctoral experience and would, therefore, possess a sophisticated understanding of the molecular biology of nucleic acids and their synthesis.

The '551 application explicitly defines "modified bases" as including, *inter alia*, "methylated bases; bisulfite-converted bases" that "may exhibit different base pairing

and/or base stacking properties as compared to native bases." Ex. 2026, ¶ [0021]. Moreover, the '551 application teaches that "uracil can be found in DNA as the result of … of bisulfite-conversion of cytosine in a common protocol used to discriminate methylated cytosine through DNA sequencing." *Id.*, ¶ [0023]. We consequently find that one of ordinary skill in the art would have known that treating DNA with bisulfite, according to a protocol well-known in the art, would result in the conversion of cytosine (but not 5-methylcytosine) to uracil, which pairs, according to the Crick-Watson rules, preferentially with adenine rather than guanine. *See, e.g.*, Levy Decl. 1, ¶ 21.

The '551 application also teaches that "redundant sequence information can be generated by sequencing the same template nucleic acid multiple times by, e.g., sequencing the same template nucleic acid molecule repeatedly or by sequencing multiple copies of the same template nucleic acid" and that:

> Such redundant sequence information can provide additional information for discriminating modified from unmodified nucleotides. For example, sequence reads from the sense or "forward" strand can be compared to sequence reads from the antisense or "reverse" strand for the same nucleic acid template to further validate the existence of one or more modified bases in the template nucleic acid. In some preferred applications, the forward and reverse strands are sequenced in a single template nucleic acid.

Ex. 2026, ¶ [0017]. The '551 application also teaches:

> Sufficiently redundant sequence data would facilitate detection of such a change in misincorporation rate. In particular, uracil (U) has a propensity to base pair with adenine (A) and somewhat with guanine (G), so with redundant sequencing data and knowledge of the misincorporation frequencies, the presence of U can be detected in a template nucleic acid. For example, both A and G would be incorporated at an "unmethylated" C site (bisulfite-converted to U) while only G would be incorporated at a C site that was not bisulfite-converted to U

Id. We consequently find that the '551 application teaches the portion of limitation c of the count that recites: "determining the sequence of the double stranded nucleic acid sample and the position of the at least one modified base in the sequence of the double stranded nucleic acid sample by comparing the sequences of the forward and reverse strands of the circular pair-locked molecule." The '551 application teaches "determining the sequence of the double stranded nucleic acid sample" by teaching "sequence reads from the … 'forward' strand can be compared to sequence reads from the … 'reverse strand for the same nucleic acid template."

ITRI's declarant, Dr. Levy, opines that "the '551 application discloses different ways of determining a modified base position that do not use a mismatch between two strands." Levy Decl. 1, ¶ 33. According to Dr. Levy, "determining a modified base position in methods of the '551 application must rely on a distinct signal from the sequencing of the modified base itself, that does not occur with an unmodified base one wishes to distinguish it from." *Id.* Dr. Levy contends that the passages from paragraph 17 of the '551 application cited *supra* "does not describe using a mismatch in a sample or CPLM involving a modified base to identify a modified base's position." *Id.*, ¶ 36.

Moreover, Dr. Levy opines that the '551 application also does not inherently describe determining the position of a modified base by using a mismatch involving a modified base in a sample or CPLM. *Id.*, ¶ 39. According to Dr. Levy, if there were a mismatched modified base, the information from the forward and reverse sequences would not be redundant, because it would allow one to determine that the strands disagree—something that would not be possible from sequencing either strand alone. *Id.*, ¶ 42. Dr. Levy states further that:

> One of ordinary skill would not understand the sentence about comparing forward and reverse strands to necessarily involve a non-redundant disagreement (revealing a mismatch) because it begins with the words "For example" and follows three sentences in a row that each discuss "redundant sequence information." Ex. 2026, ¶ 17. Thus, the '551 application indicates that comparing the sequences of the forward and reverse strands is a way to use the redundancy discussed in the previous sentences.

Id., ¶43. Dr. Levy concludes:

> [T]he mention of comparing strands in the '551 application would not be understood as requiring a mismatch or disclosing the use of a mismatch to identify a modified base position. Instead, a consensus sequence could be generated for a molecule that contains a modified base such as 5-methylcytosine. One could do so by making and sequencing a circular molecule, and then comparing forward and reverse strand sequences according to the example in the '696 application. A match, i.e., redundant sequence information, involving the modified base (e.g., resulting from a pairing of the 5-methylcytosine to guanine) would aid in constructing the consensus sequence, because the corresponding forward and reverse strand sequence data agree.

Id., ¶ 52.

We are not convinced. The '551 application teaches that "sequence reads from the … "forward" strand can be compared to sequence reads from the … "reverse" strand for the same nucleic acid template to further validate the existence of one or more modified bases in the template nucleic acid. Ex. 2026, ¶ [0017]. The '551 application thus teaches comparing the sequences of forward and reverse strands. ITRI argues that this teaches for detecting matched base pairs, but not mismatched base pairs; Dr. Levy specifically states that a match resulting from the pairing of *unconverted* 5-methylcytosine with guanine, would aid in constructing the consensus sequence, because the forward and reverse strand data would agree. Levy Decl. 1, ¶ 52.

However, given the relatively high level of skill of the average artisan, as we have defined it, it is not reasonable to believe that such a person would not recognize a mismatch occurring between a modified base, i.e., cytosine converted to uracil by bisulfite treatment (as taught by the '551 application) paired with guanine, and not understand from that mismatch that "at least one modified base in the double-stranded nucleic sample is paired with a base having a base pairing specificity different from its preferred partner base" as required by the disputed limitation. To reason otherwise, to contend that a person of ordinary skill would only look at "correct" base pair matches and discard, without thought, the mismatched pairs, is not reasonable because the person of ordinary skill would know, *a priori*, that bisulfite-converted uracil would be expected to be paired, incorrectly, with the guanine that would normally be paired with the pre-converted cytosine. In other words, a person of ordinary skill would expect to see mismatched uracil-guanine base pairs in a DNA sample strand that had been modified by bisulfite-conversion. Ex. 2026, ¶ [0017]

The '551 application teaches, by way of example, "sequence reads from the sense or "forward" strand can be compared to sequence reads from the antisense or "reverse" strand for the same nucleic acid template to further validate the existence of one or more modified bases in the template nucleic acid." Ex. 2026, ¶ [0017]. We do not agree with ITRI that the use of "validate" in this passage of the '551 application necessarily requires that the analysis of base pair only validate that the base pairs agree with the Crick-Watson rules. To the contrary, we find that finding a mismatched uracil (or thymine)-guanine pair would "validate", commonly understood to mean "support or corroborate," the existence of a modified base (i.e., a bisulfite-modified cytosine) in the template

nucleic acid on a sound or authoritative basis (*viz.*, knowledge of the bisulfite conversion reaction).

ITRI has not shown that the ’551 application, when viewed in light of the ’630 Specification, fails to disclose “comparing forward and reverse strands” “to further validate the existence of one or more modified bases in the template nucleic acid.” We consequently determine that ITRI has not met its burden of showing why we should deny PacBio the benefit accorded its’551 application. ITRI’s Motion 2 is DENIED.

V. Conclusion

For the reasons set forth above, PacBio’s Motion 1 to rescind the benefit accorded to ITRI for its’3l3 application for failure to provide a constructive reduction to practice of Count 1 is DENIED. PacBio’s Motion 2 seeking judgment of unpatentability under 35 U.S.C. § 103(a) against ITRI’s ’630 patent is GRANTED. ITRI’S Motion 2 to deny PacBio the benefit of its ’551 application is DENIED.

Attorney for Pan:

David S. Forman, Ph.D.
M. Paul Barker
Adam M. Breier, Ph.D.
Finnegan, Henderson, Farabow, Garrett & Dunner, LLP
David.forman@finnegan.com
Paul.barker@finnegan.com
Adam.breier@finnegan.com

Attorney for Flusberg:

Michael S. Tuscan, Ph.D.
Bonnie Weiss McLeod, Ph.D.
Cooley, LLP
mtuscan@cooley.com
bweissmcleod@cooley.com
zpatdocketing@cooley.com

(12) **United States Patent**
Pan et al.

(10) **Patent No.: US 8,486,630 B2**
(45) **Date of Patent: Jul. 16, 2013**

(54) **METHODS FOR ACCURATE SEQUENCE DATA AND MODIFIED BASE POSITION DETERMINATION**

(75) Inventors: **Chao-Chi Pan**, Hsinchu (TW); **Jenn-Yeh Fann**, Hsinchu County (TW); **Chung-Fan Chiou**, Hsinchu County (TW); **Hung-Chi Chien**, Hsinchu (TW); **Hui-Ling Chen**, Hsinchu (TW)

(73) Assignee: **Industrial Technology Research Institute**, Hsinchu (TW)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 643 days.

(21) Appl. No.: **12/613,291**

(22) Filed: **Nov. 5, 2009**

(65) **Prior Publication Data**

US 2010/0121582 A1 May 13, 2010

Related U.S. Application Data

(60) Provisional application No. 61/112,548, filed on Nov. 7, 2008, provisional application No. 61/167,313, filed on Apr. 7, 2009.

(51) **Int. Cl.**
C12Q 1/68 (2006.01)

(52) **U.S. Cl.**
USPC **435/6.11**; 435/6.12

(58) **Field of Classification Search**
None
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

7,282,337	B1	10/2007	Harris
7,399,614	B2	7/2008	Zon
8,153,375	B2 *	4/2012	Travers et al.
2002/0182630	A1	12/2002	Milosavljevic
2005/0142559	A1	6/2005	Makrigiorgos
2006/0024711	A1	2/2006	Lapidus et al.
2007/0231808	A1	10/2007	Gouda et al.
2009/0280538	A1	11/2009	Patel et al.
2009/0298075	A1	12/2009	Travers et al.
2010/0221716	A1	9/2010	Flusberg et al.

FOREIGN PATENT DOCUMENTS

WO	WO 2009/017678	2/2009

OTHER PUBLICATIONS

Adessi et al., "Solid Phase DNA Amplification: Characterisation of Primer Attachment and Amplification Mechanisms", Nucleic Acids Research, vol. 28, No. 20 e87, pp. 1-8 (2000).
Branton et al., "The Potential and Challenges of Nanopore Sequencing", Nat Biotechnol., vol. 26, pp. 1146-1153 (Oct. 2008).
Braslavsky et al., "Sequence Information Can Be Obtained From Single DNA Molecules", PNAS, vol. 100, pp. 3960-3964 (2003).
Brenner et al., "Gene Expression Analysis by Massively Parallel Signature Sequencing (MPSS) on Microbead Arrays", Nature Biotechnology, vol. 18, pp. 630-634 (2000).
Brown, "Mutation, Repair, and Recombination", Genomes Second Edition, Chapter 14, pp. 418-457 (2002), John Wiley & Sons, New York.
Chen et al., "DNA Cleavage in Trans by the Active Site Tyrosine During FLP Recombination: Switching Protein Partners Before Exchanging Strands", Cell, vol. 69, pp. 647-658 (1992).
Eid et al. "Real-Time DNA Sequencing From Single Polymerase Molecules", Science, vol. 323, pp. 133-138 (2009).
Guo et al., "Weakening of the T7 Promoter-Polymerase Interaction Facilitates Promoter Release", The Journal of Biological Chemistry, vol. 280, pp. 14956-14961 (2005).
Harris et al., "Single-Molecule DNA Sequencing of a Viral Genome", Science, vol. 320, pp. 106-109 (Apr. 2008).
Kato, "Impact of the Next Generation DNA Sequencers", Int J Clin Exp Med, vol. 2, pp. 193-202 (Jul. 2009).
Korlach et al., "Selective Aluminum Passivation for Targeted Immobilization of Single DNA Polymerase Molecules in Zero-Mode Waveguide Nanostructures", Proc. Natl. Acad. Sci. USA, vol. 105, pp. 1176-1181 (Jan. 2008).
Laird et al., "Hairpin-Bisulfite PCR: Assessing Epigenetic Methylation Patterns on Complementary Strands of Individual DNA Molecules", Proc. Natl. Acad. Sci. USA, vol. 101, pp. 204-209 (2004).
Levene et al., "Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations", Science, vol. 299, pp. 682-686 (2003).
Margulies et al., "Genome Sequencing in Microfabricated High-Density Picolitre Reactors", Nature, vol. 437, pp. 376-380 (2005); corrigendum at vol. 441, p. 120 (2006).
Matsumura et al., "Photochemical Transition of 5-Methylcytosine to Thymine by DNA Photoligation", Nucleic Acids Symposium Series No. 51, pp. 233-234 (2007).
Ronaghi et al., "Real-Time DNA Sequencing Using Detection of Pyrophosphate Release", Analytical Biochemistry, Vo. 242, pp. 84-89 (1996).
Roth et al., "Characterization of Broken DNA Molecules Assoicated With V(D)J Recombination", Proc. Natl. Acad. Sci. USA, vol. 90, pp. 10788-10792 (1993).
Ruparel et al., "Design and Synthesis of a 3-O-Allyl Photocleavable Fluorescent Nucleotide as a Reversible Terminator for DNA Sequencing by Synthesis", Proc. Natl. Acad. Sci. USA 102, 5932-5937 (2005).
Smith et al., "Rapid Whole-Genome Mutational Profiling Using Next-Generation Sequencing Technologies", Genome Research, vol. 18, pp. 1638-1642 (2009).
Zilberman et al., "Genome-Wide Analysis of DNA Methylation Patterns", Development, vol. 134, pp. 3959-3965 (2007).

(Continued)

Primary Examiner — James Martinell
(74) *Attorney, Agent, or Firm* — Finnegan, Henderson, Farabow, Garrett & Dunner, L.L.P.

(57) **ABSTRACT**

Disclosed herein are methods of determining the sequence and/or positions of modified bases in a nucleic acid sample present in a circular molecule with a nucleic acid insert of known sequence comprising obtaining sequence data of at least two insert-sample units. In some embodiments, the methods comprise obtaining sequence data using circular pair-locked molecules. In some embodiments, the methods comprise calculating scores of sequences of the nucleic acid inserts by comparing the sequences to the known sequence of the nucleic acid insert, and accepting or rejecting repeats of the sequence of the nucleic acid sample according to the scores of one or both of the sequences of the inserts immediately upstream or downstream of the repeats of the sequence of the nucleic acid sample.

28 Claims, 17 Drawing Sheets

OTHER PUBLICATIONS

International Search Report and Written Opinion mailed Feb. 25, 2010, for counterpart International Application No. PCT/CN2009/074851.

Pacific Biosciences, "Harnessing Nature's Powerful Sequencing Engine: Single-Molecule Real-Time DNA Sequencing," slides 1-12 and 27-36, labeled Cold Spring Harbor Personal Genomes Meeting, Oct. 12, 2008.

Author Unknown, Cold Spring Harbor Personal Genomes Meeting Program, pp. 1-4, Oct. 2008.

* cited by examiner

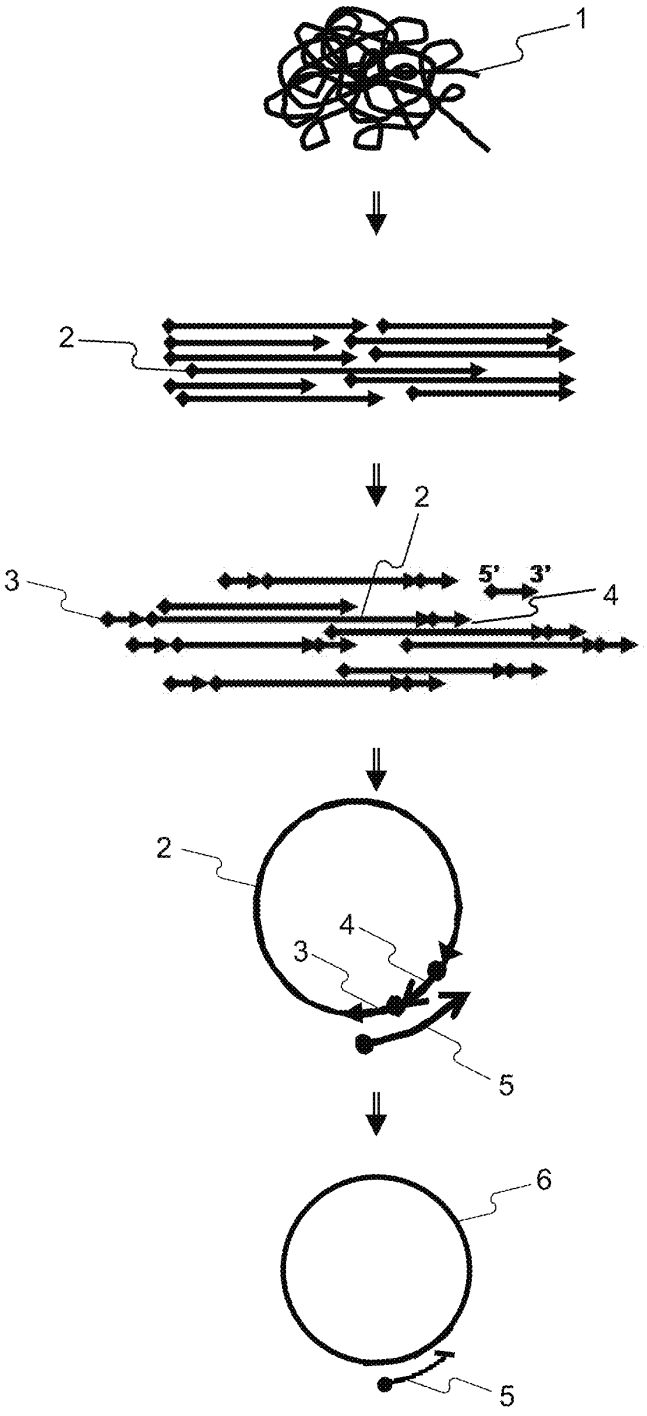


Fig. 1

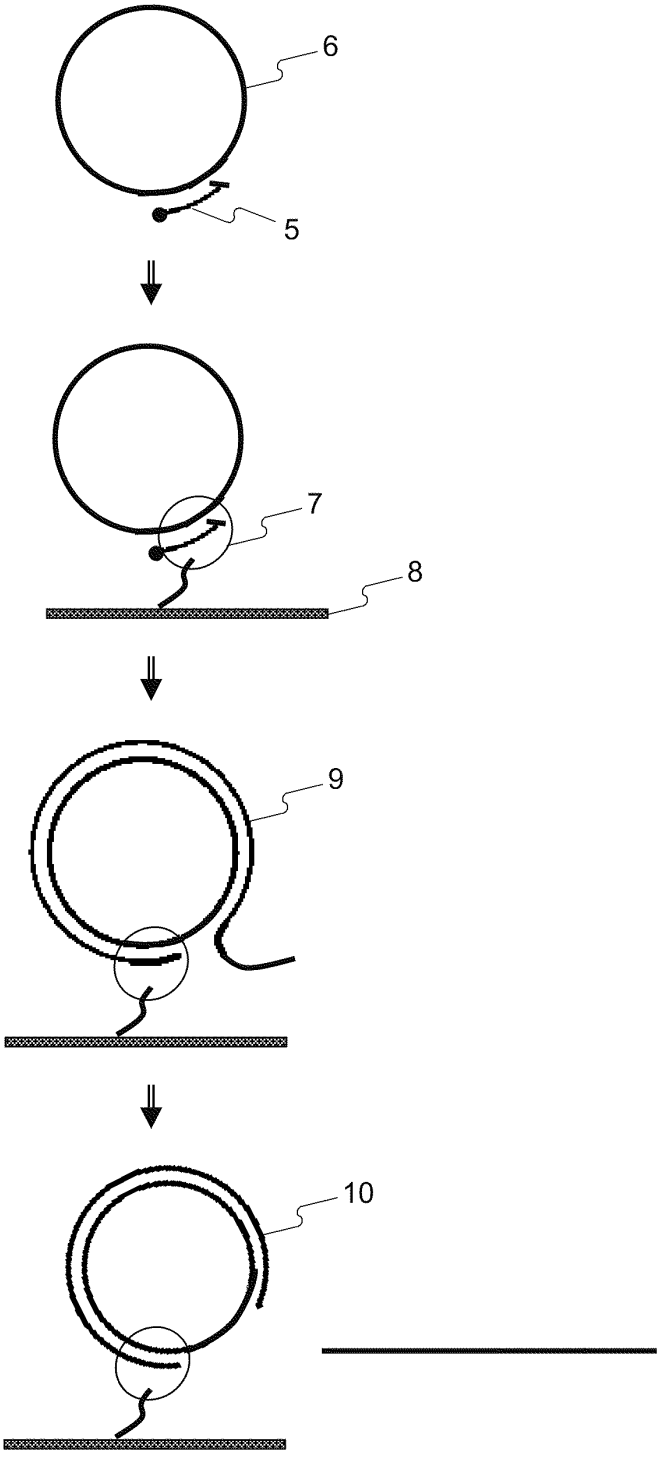


Fig. 2

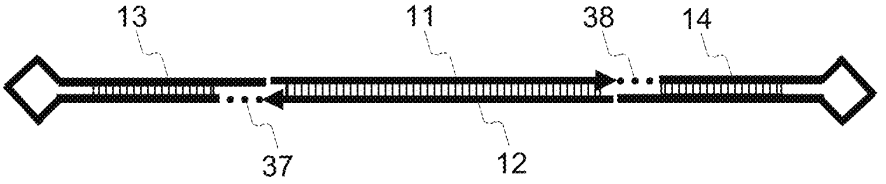

13
11
38
14
37
12

Fig. 3A

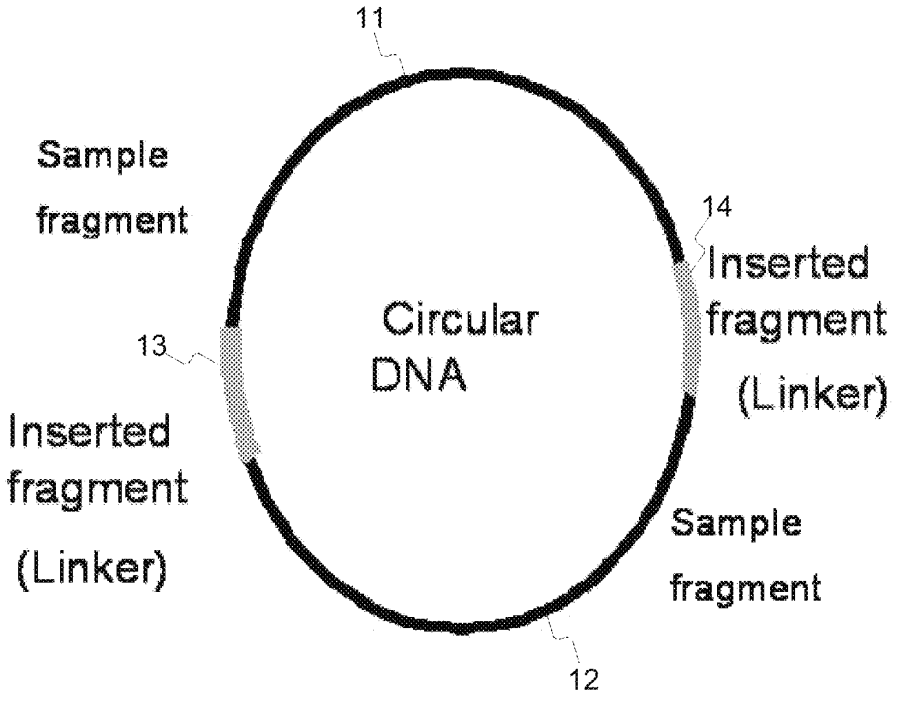

11
Sample fragment
14
Inserted fragment (Linker)
Circular DNA
13
Inserted fragment (Linker)
Sample fragment
12

Fig. 3B

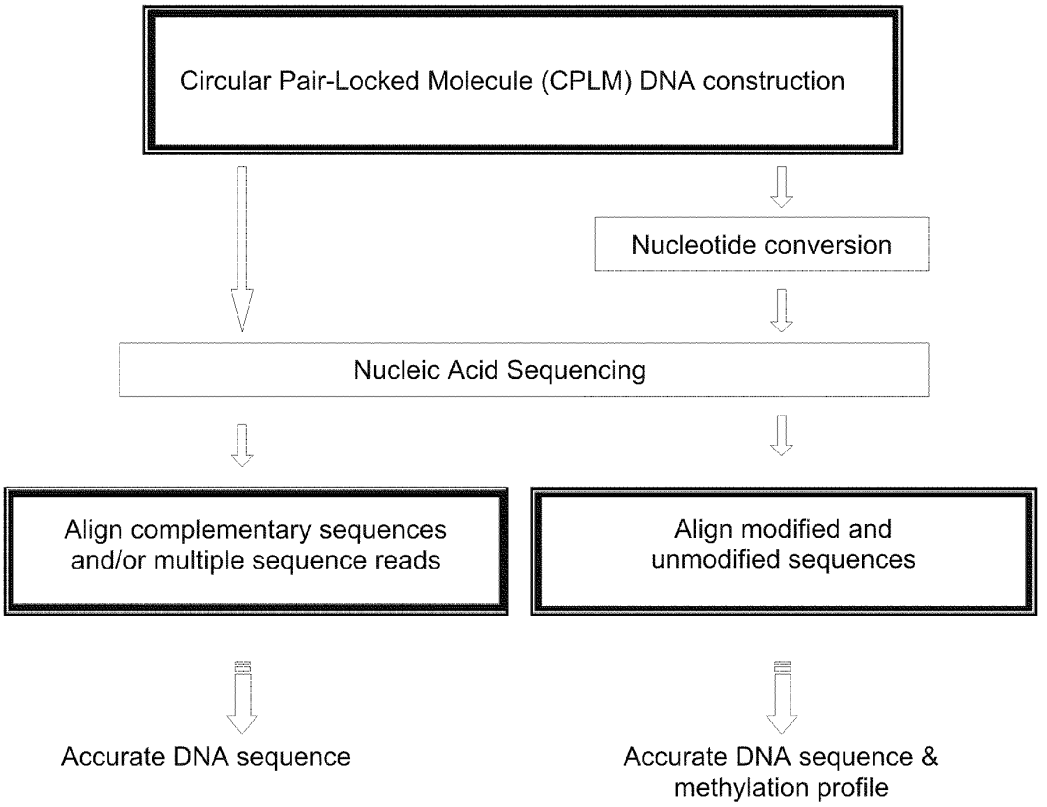

Circular Pair-Locked Molecule (CPLM) DNA construction
Nucleotide conversion
Nucleic Acid Sequencing
Align complementary sequences and/or multiple sequence reads
Align modified and unmodified sequences
Accurate DNA sequence
Accurate DNA sequence & methylation profile

Fig. 4

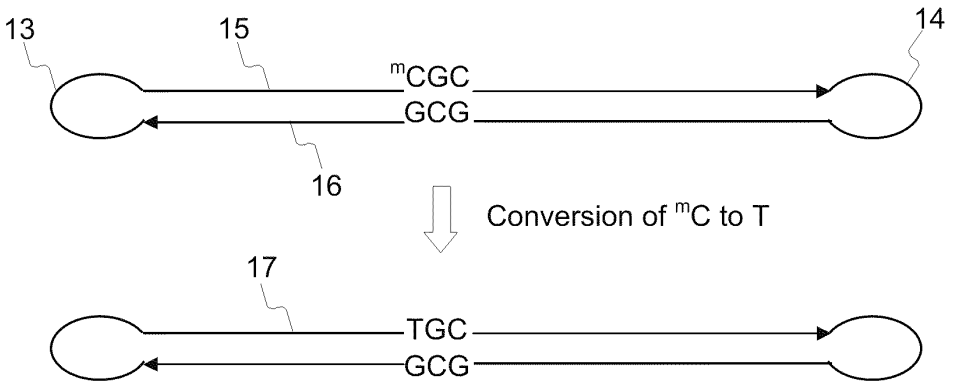

13
15
14
mCGC
GCG
16
Conversion of mC to T
17
TGC
GCG

Fig. 5A

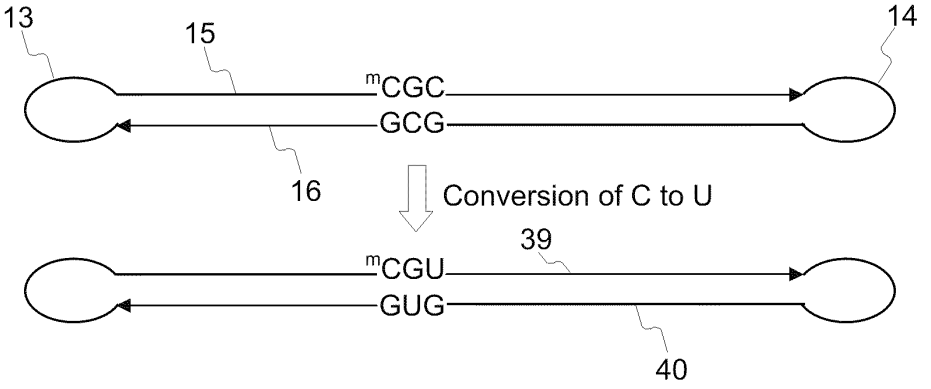

13
15
14
mCGC
GCG
16
Conversion of C to U
39
mCGU
GUG
40

Fig. 5B

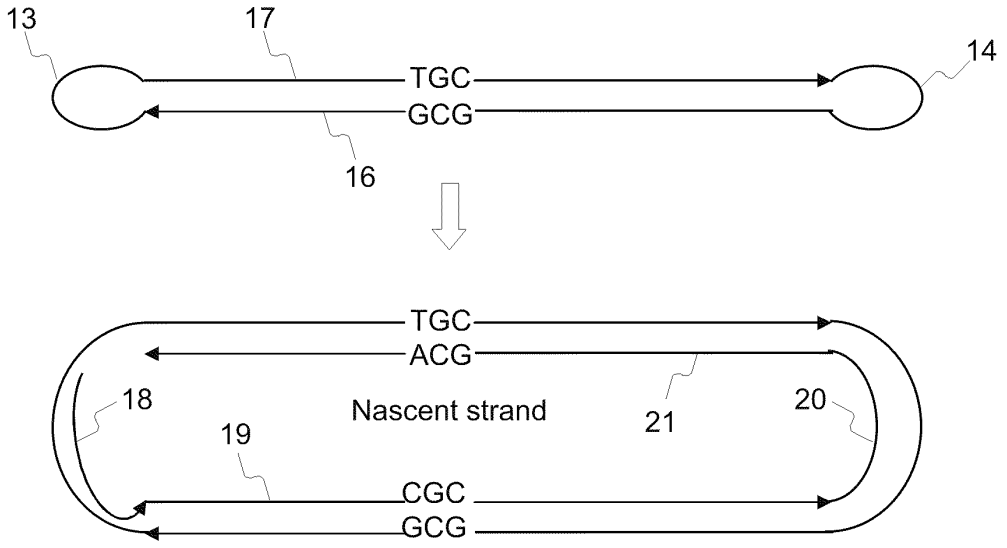

13
17
14
TGC
GCG
16
TGC
ACG
18
Nascent strand
21
20
19
CGC
GCG

Fig. 6A

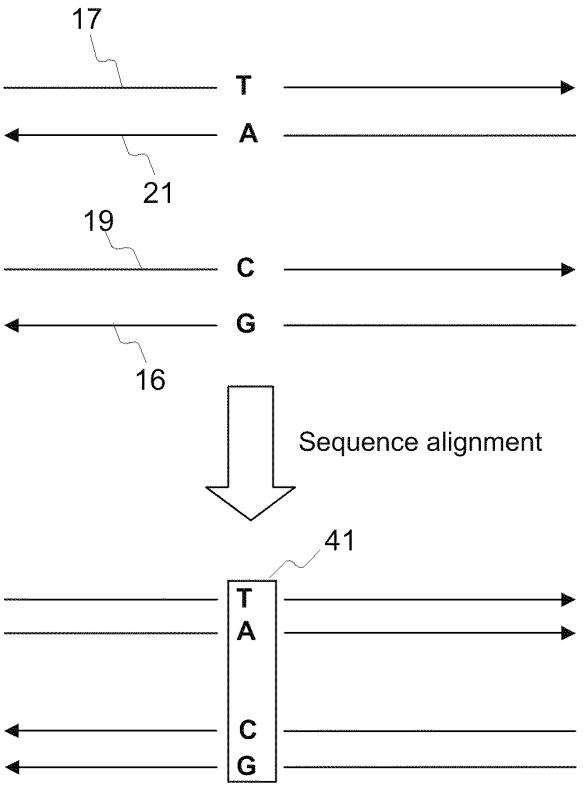

17
T
A
21
19
C
G
16
Sequence alignment
41
T
A
C
G

Fig. 6B

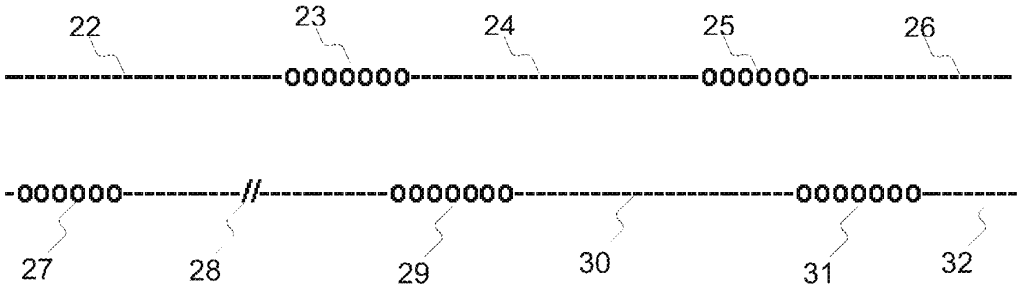

22
23
24
25
26
27
28
29
30
31
32
Reading result from a circular template

Fig. 7A

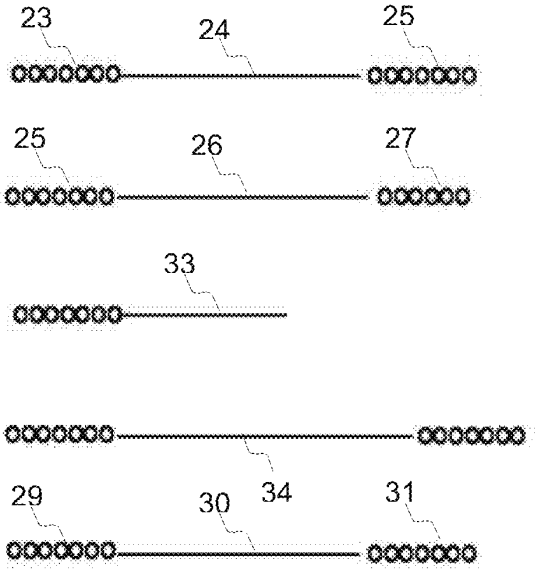

23
24
25
25
26
27
33
34
29
30
31

Fig. 7B

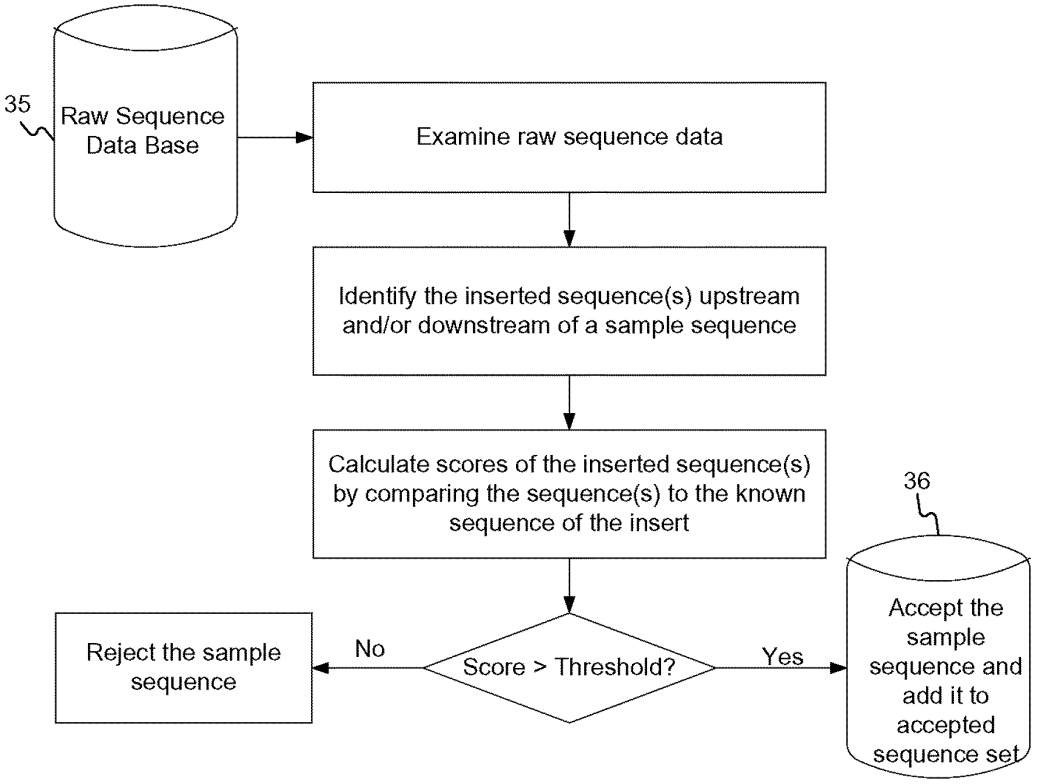

35
Raw Sequence Data Base
Examine raw sequence data
Identify the inserted sequence(s) upstream and/or downstream of a sample sequence
Calculate scores of the inserted sequence(s) by comparing the sequence(s) to the known sequence of the insert
36
Accept the sample sequence and add it to accepted sequence set
Yes
Score > Threshold?
No
Reject the sample sequence

Fig. 8

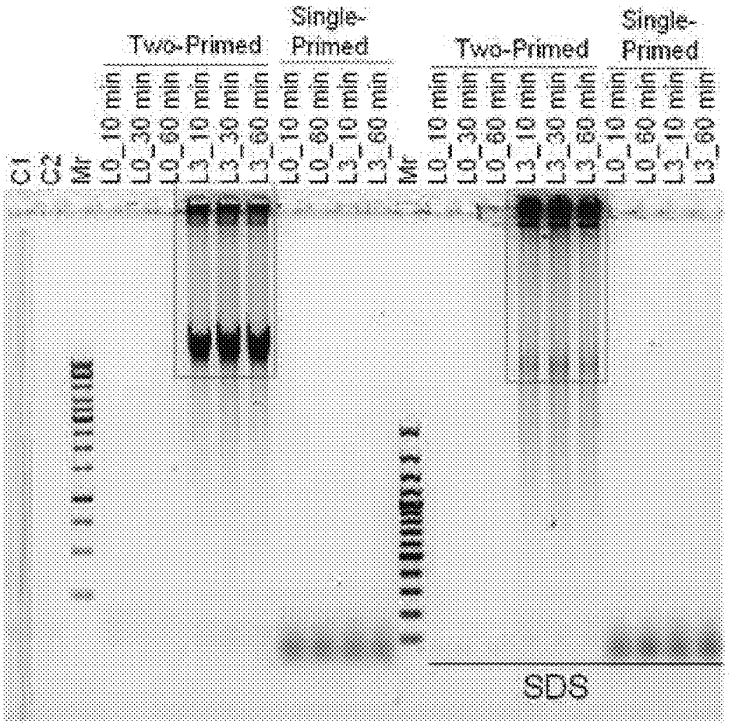

Two-Primed
Single-Primed
Two-Primed
Single-Primed
C1
C2
Mr
L0_10 min
L0_30 min
L0_60 min
L3_10 min
L3_30 min
L3_60 min
L0_10 min
L0_60 min
L3_10 min
L3_60 min
Mr
L0_10 min
L0_30 min
L0_60 min
L3_10 min
L3_30 min
L3_60 min
L0_10 min
L0_60 min
L3_10 min
L3_60 min
SDS

Fig. 9

Fig. 10A

Fig. 10B

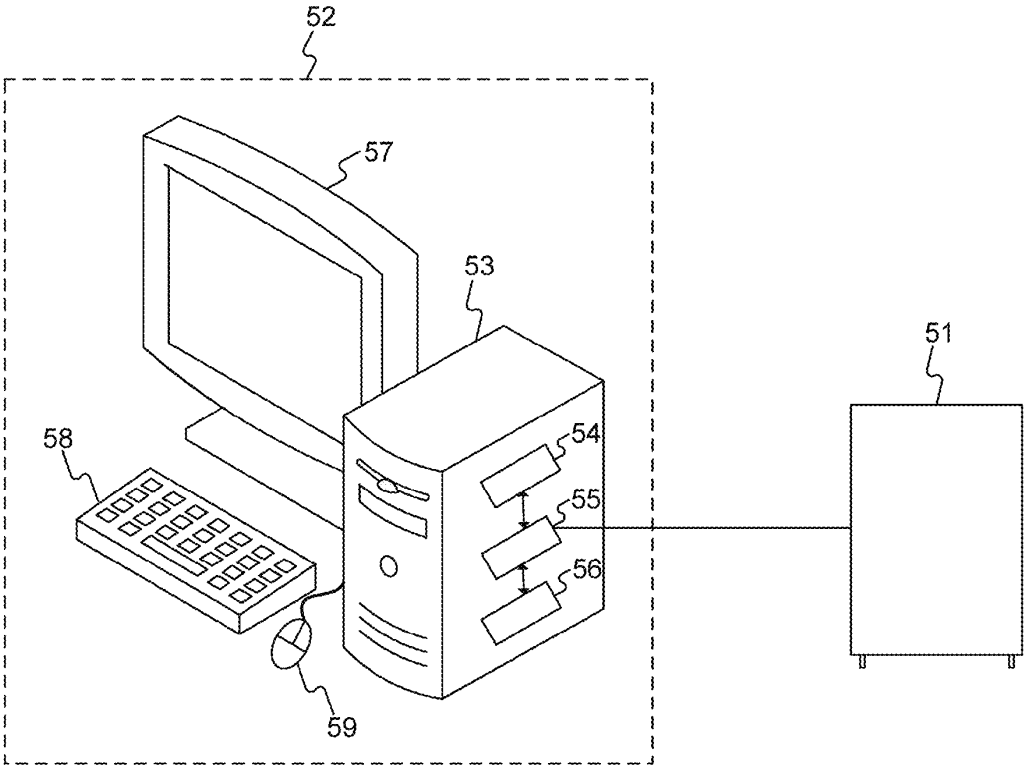


Fig. 11A

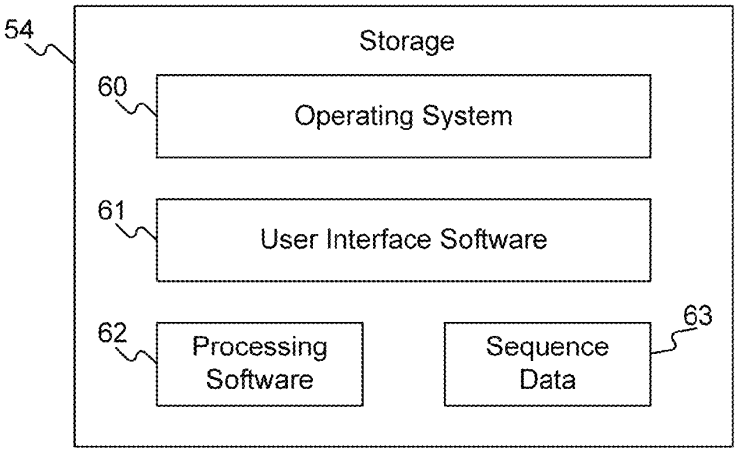


Fig. 11B

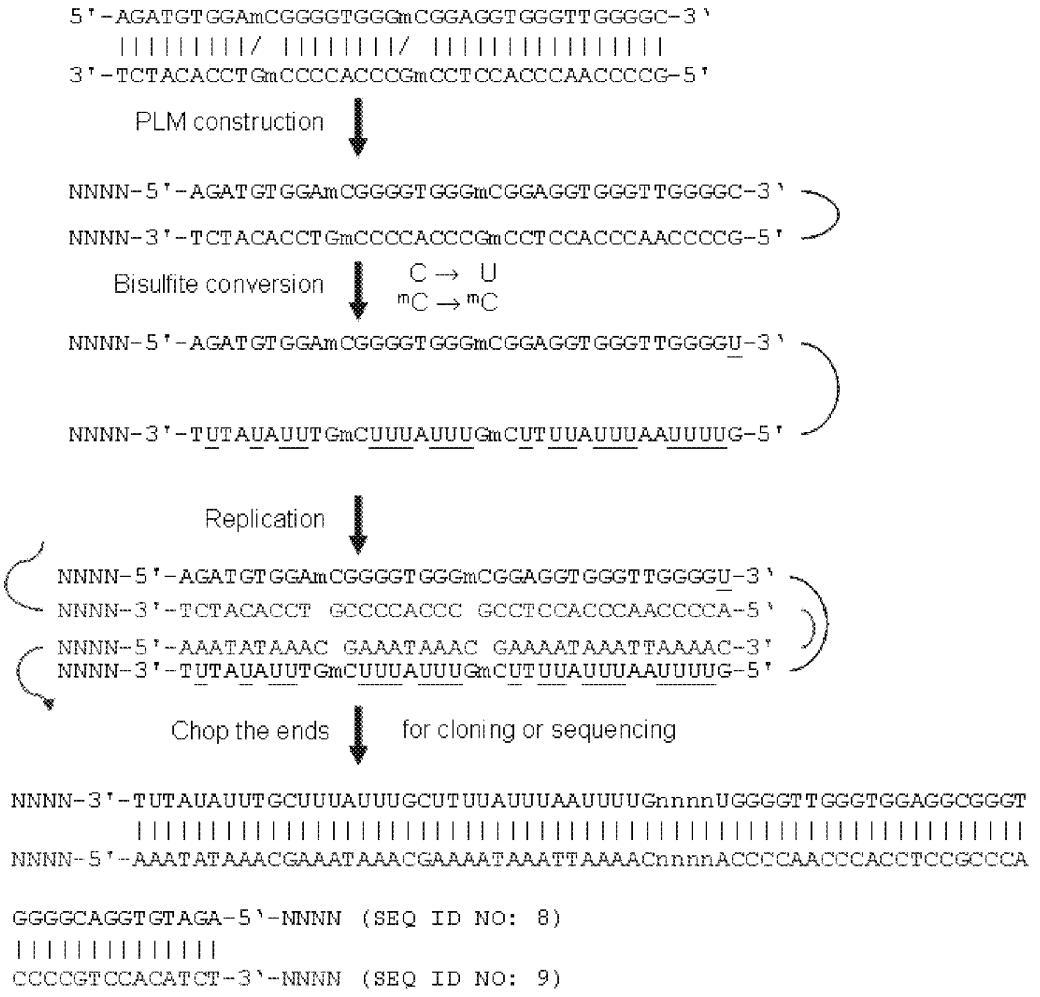

5'-AGATGTGGAmCGGGGTGGGmCGGAGGTGGGTTGGGGC-3'
3'-TCTACACCTGmCCCCACCCGmCCTCCACCCAACCCCG-5'
PLM construction
NNNN-5'-AGATGTGGAmCGGGGTGGGmCGGAGGTGGGTTGGGGC-3'
NNNN-3'-TCTACACCTGmCCCCACCCGmCCTCCACCCAACCCCG-5'
Bisulfite conversion
C → U
mC → mC
NNNN-5'-AGATGTGGAmCGGGGTGGGmCGGAGGTGGGTTGGGGU-3'
NNNN-3'-TUTAUAUUTGmCUUUAUUUGmCUTUUAUUUAAUUUUG-5'
Replication
NNNN-5'-AGATGTGGAmCGGGGTGGGmCGGAGGTGGGTTGGGGU-3'
NNNN-3'-TCTACACCT GCCCCACCC GCCTCCACCCAACCCCA-5'
NNNN-5'-AAATATAAAC GAAATAAAC GAAAATAAATTAAAAC-3'
NNNN-3'-TUTAUAUUTGmCUUUAUUUGmCUTUUAUUUAAUUUUG-5'
Chop the ends
for cloning or sequencing
NNNN-3'-TUTAUAUUTGCUUUAUUUGCUTUUAUUUAAUUUUGnnnnUGGGGTTGGGTGGAGGCGGGT
NNNN-5'-AAATATAAACGAAATAAACGAAAATAAATTAAAACnnnnACCCCAACCCACCTCCGCCCA
GGGGCAGGTGTAGA-5'-NNNN (SEQ ID NO: 8)
CCCCGTCCACATCT-3'-NNNN (SEQ ID NO: 9)

Fig. 12

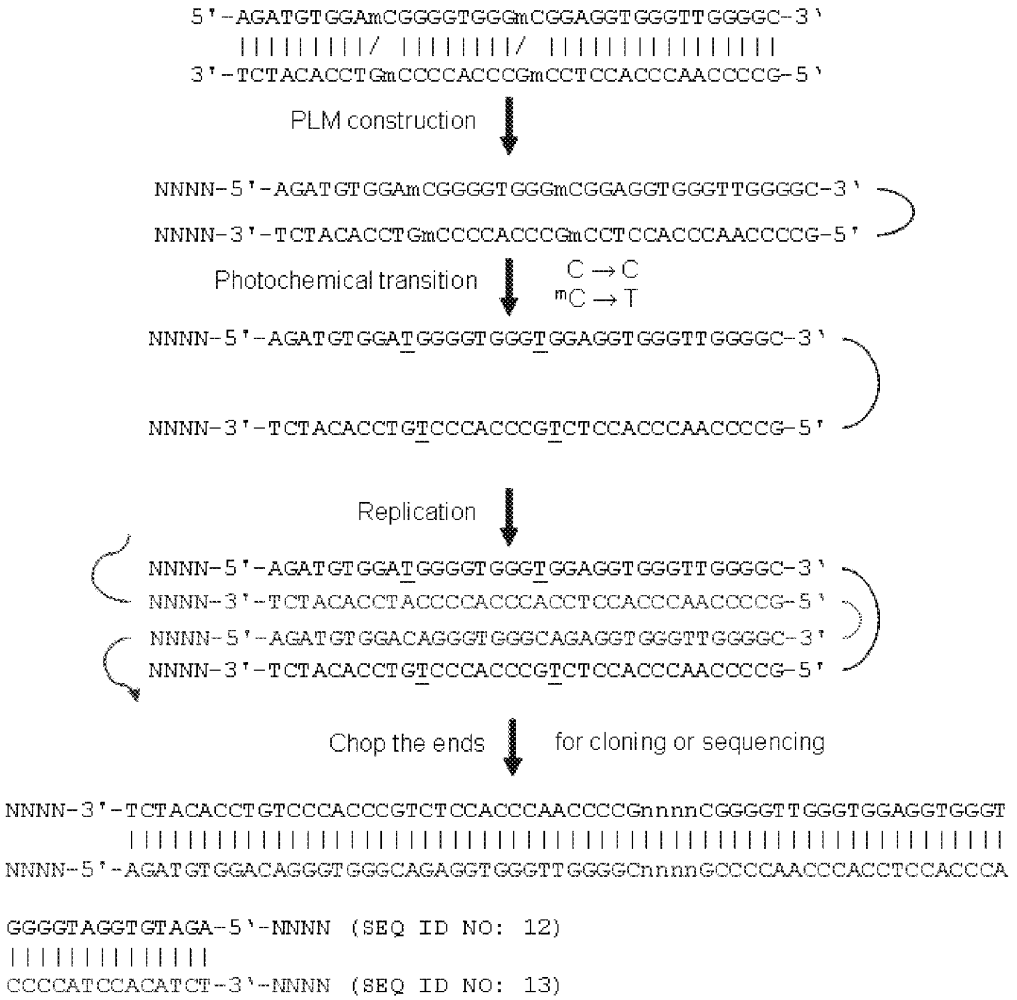

5'-AGATGTGGAmCGGGGTGGGmCGGAGGTGGGTTGGGGC-3'
3'-TCTACACCTGmCCCCACCCGmCCTCCACCCAACCCCG-5'
PLM construction
NNNN-5'-AGATGTGGAmCGGGGTGGGmCGGAGGTGGGTTGGGGC-3'
NNNN-3'-TCTACACCTGmCCCCACCCGmCCTCCACCCAACCCCG-5'
Photochemical transition
C → C
mC → T
NNNN-5'-AGATGTGGATGGGGTGGGTGGAGGTGGGTTGGGGC-3'
NNNN-3'-TCTACACCTGTCCCACCCGTCTCCACCCAACCCCG-5'
Replication
NNNN-5'-AGATGTGGATGGGGTGGGTGGAGGTGGGTTGGGGC-3'
NNNN-3'-TCTACACCTACCCCACCCACCTCCACCCAACCCCG-5'
NNNN-5'-AGATGTGGACAGGGTGGGCAGAGGTGGGTTGGGGC-3'
NNNN-3'-TCTACACCTGTCCCACCCGTCTCCACCCAACCCCG-5'
Chop the ends
for cloning or sequencing
NNNN-3'-TCTACACCTGTCCCACCCGTCTCCACCCAACCCCGnnnnCGGGGTTGGGTGGAGGTGGGT
NNNN-5'-AGATGTGGACAGGGTGGGCAGAGGTGGGTTGGGGCnnnnGCCCCAACCCACCTCCACCCA
GGGGTAGGTGTAGA-5'-NNNN (SEQ ID NO: 12)
CCCCATCCACATCT-3'-NNNN (SEQ ID NO: 13)

Fig. 13

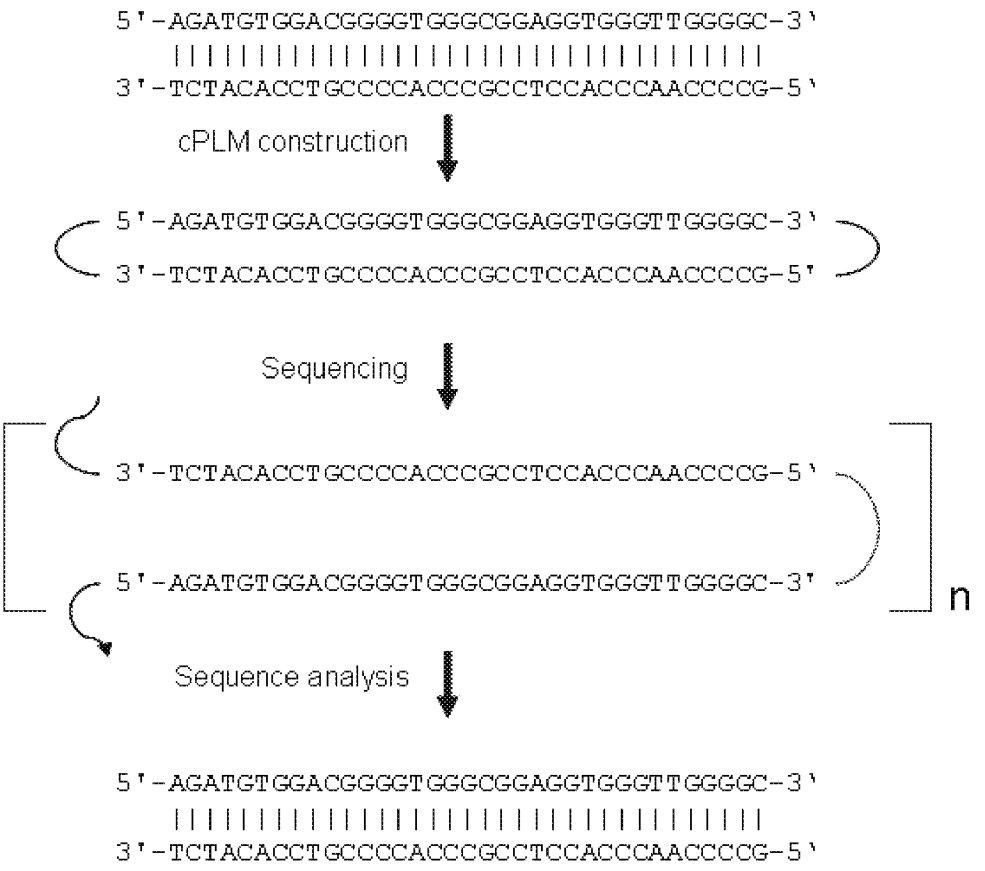

5'-AGATGTGGACGGGGTGGGCGGAGGTGGGTTGGGGC-3'
3'-TCTACACCTGCCCCACCCGCCTCCACCCAACCCCG-5'
cPLM construction
5'-AGATGTGGACGGGGTGGGCGGAGGTGGGTTGGGGC-3'
3'-TCTACACCTGCCCCACCCGCCTCCACCCAACCCCG-5'
Sequencing
3'-TCTACACCTGCCCCACCCGCCTCCACCCAACCCCG-5'
5'-AGATGTGGACGGGGTGGGCGGAGGTGGGTTGGGGC-3'
n
Sequence analysis
5'-AGATGTGGACGGGGTGGGCGGAGGTGGGTTGGGGC-3'
3'-TCTACACCTGCCCCACCCGCCTCCACCCAACCCCG-5'

Fig. 14

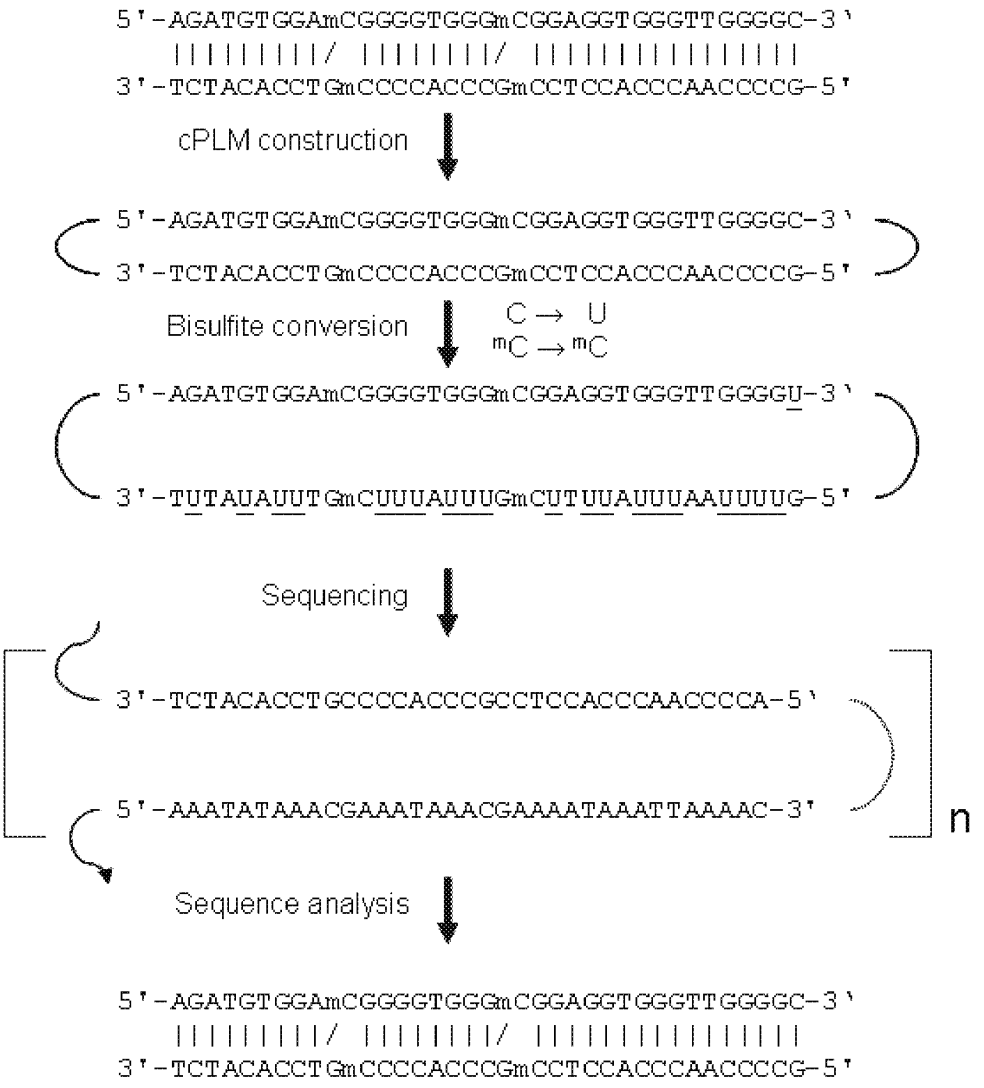

5'-AGATGTGGAmCGGGGTGGGmCGGAGGTGGGTTGGGGC-3'
3'-TCTACACCTGmCCCCACCCGmCCTCCACCCAACCCCG-5'
cPLM construction
5'-AGATGTGGAmCGGGGTGGGmCGGAGGTGGGTTGGGGC-3'
3'-TCTACACCTGmCCCCACCCGmCCTCCACCCAACCCCG-5'
Bisulfite conversion
C → U
mC → mC
5'-AGATGTGGAmCGGGGTGGGmCGGAGGTGGGTTGGGGU-3'
3'-TUTAUAUUTGmCUUUAUUUGmCUTUUAUUUAAUUUUG-5'
Sequencing
3'-TCTACACCTGCCCCACCCGCCTCCACCCAACCCCA-5'
5'-AAATATAAACGAAATAAACGAAAATAAATTAAAAC-3'
n
Sequence analysis
5'-AGATGTGGAmCGGGGTGGGmCGGAGGTGGGTTGGGGC-3'
3'-TCTACACCTGmCCCCACCCGmCCTCCACCCAACCCCG-5'

Fig. 15

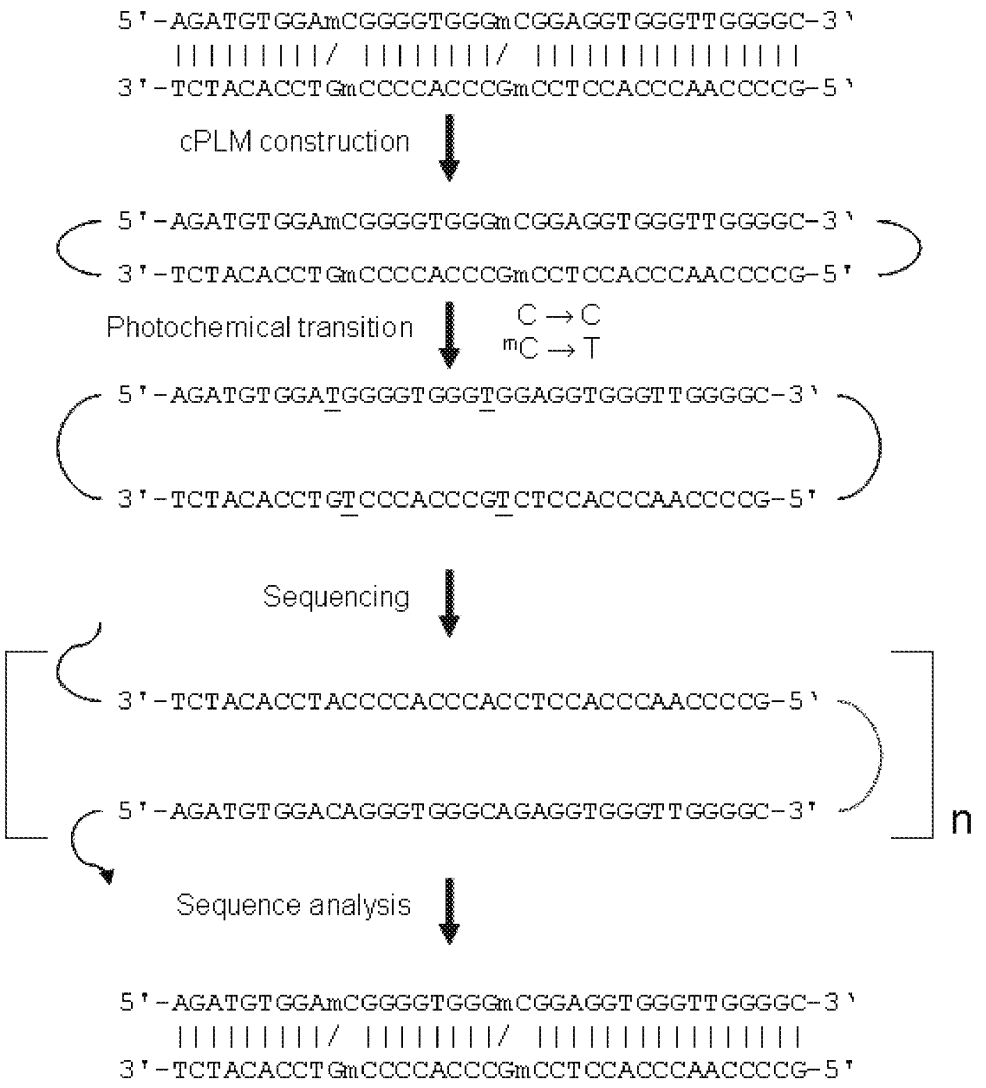

5'-AGATGTGGAmCGGGGTGGGmCGGAGGTGGGTTGGGGC-3'
3'-TCTACACCTGmCCCCACCCGmCCTCCACCCAACCCCG-5'
cPLM construction
5'-AGATGTGGAmCGGGGTGGGmCGGAGGTGGGTTGGGGC-3'
3'-TCTACACCTGmCCCCACCCGmCCTCCACCCAACCCCG-5'
Photochemical transition
C → C
mC → T
5'-AGATGTGGATGGGGTGGGTGGAGGTGGGTTGGGGC-3'
3'-TCTACACCTGTCCCACCCGTCTCCACCCAACCCCG-5'
Sequencing
3'-TCTACACCTACCCCACCCACCTCCACCCAACCCCG-5'
5'-AGATGTGGACAGGGTGGGCAGAGGTGGGTTGGGGC-3'
n
Sequence analysis
5'-AGATGTGGAmCGGGGTGGGmCGGAGGTGGGTTGGGGC-3'
3'-TCTACACCTGmCCCCACCCGmCCTCCACCCAACCCCG-5'

Fig. 16

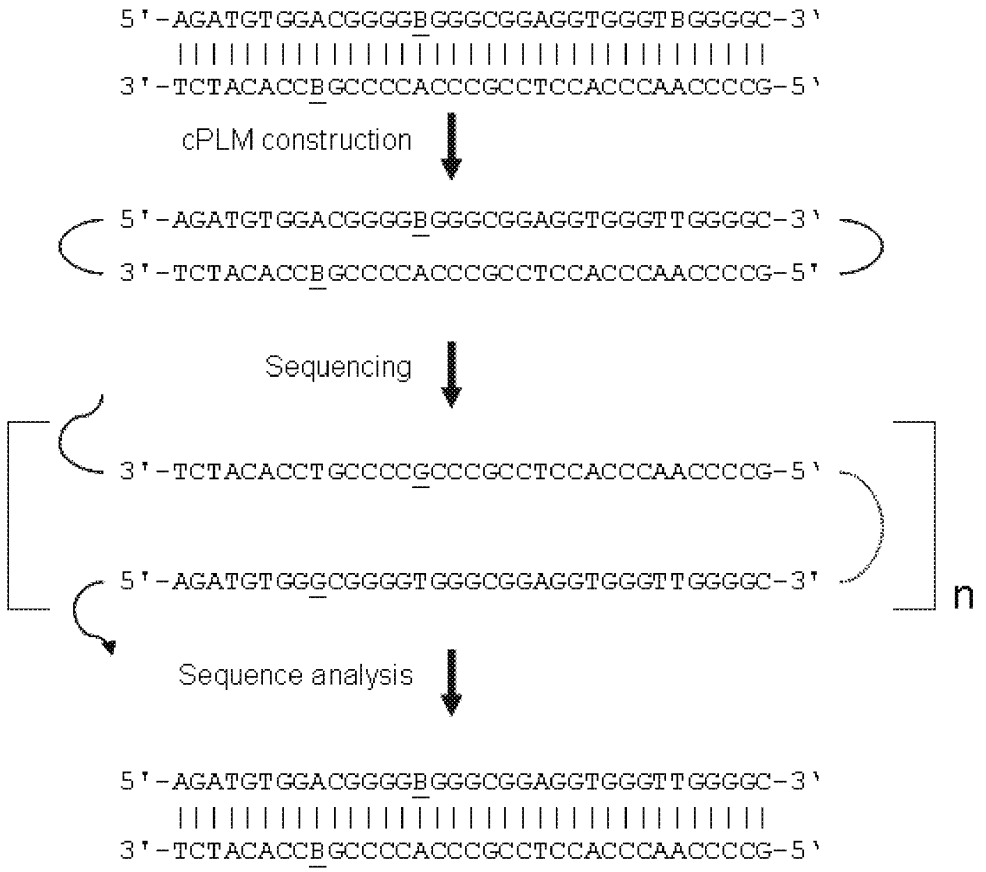

5'-AGATGTGGACGGGGBGGGCGGAGGTGGGTBGGGGC-3'
3'-TCTACACCBGCCCCACCCGCCTCCACCCAACCCCG-5'
cPLM construction
5'-AGATGTGGACGGGGBGGGCGGAGGTGGGTTGGGGC-3'
3'-TCTACACCBGCCCCACCCGCCTCCACCCAACCCCG-5'
Sequencing
3'-TCTACACCTGCCCCGCCCGCCTCCACCCAACCCCG-5'
5'-AGATGTGGGCGGGGTGGGCGGAGGTGGGTTGGGGC-3'
n
Sequence analysis
5'-AGATGTGGACGGGGBGGGCGGAGGTGGGTTGGGGC-3'
3'-TCTACACCBGCCCCACCCGCCTCCACCCAACCCCG-5'

Fig. 17

METHODS FOR ACCURATE SEQUENCE DATA AND MODIFIED BASE POSITION DETERMINATION

This application claims the benefit of U.S. Provisional Patent Application No. 61/112,548, filed on Nov. 7, 2008, and of U.S. Provisional Patent Application No. 61/167,313, filed on Apr. 7, 2009, both of which are incorporated herein by reference.

The present invention relates to methods of determining the sequence of nucleic acids and of identifying the positions of modified bases in nucleic acids.

BACKGROUND OF THE INVENTION

Recent developments in DNA sequencing technology have raised the possibility of highly personalized, preventive medicine on the genomic level. Additionally, the possibility of rapidly acquiring large amounts of sequence data from multiple individuals within one or more populations may usher in a new phase of the genomics revolution in biomedical science.

Single base differences between genotypes can have substantial phenotypic effects. For example, over 300 mutations have been identified in the gene encoding phenylalanine hydroxylase (PAH), the enzyme that converts phenylalanine to tyrosine in phenylalanine catabolism and protein and neurotransmitter biosynthesis that result in a deficient enzyme activity and lead to the disorders hyperphenylalaninaemia and phenylketonuria. See, e.g., Jennings et al., *Eur J Hum Genet* 8, 683-696 (2000).

Sequence data can be obtained using the Sanger sequencing method, in which labeled dideoxy chain terminator nucleotide analogs are incorporated in a bulk primer extension reaction and products of differing lengths are resolved and analyzed to determine the identity of the incorporated terminator. See, e.g., Sanger et al., *Proc Natl Acad Sci USA* 74, 5463-5467 (1997). Indeed, many genome sequences have been determined using this technology. However, the cost and speed of acquiring sequence data by Sanger sequencing can be limiting.

New sequencing technologies can produce sequence data at an astounding rate—hundreds of megabases per day, with costs per base lower than for Sanger sequencing. See, e.g., Kato, *Int J Clin Exp Med* 2, 193-202 (2009). However, the raw data obtained using these sequencing technologies can be more error prone than traditional Sanger sequencing. This can result from obtaining information from individual DNA molecules instead of a bulk population.

For example, in single molecule sequencing by synthesis, a base could be skipped due to the device missing a weak signal, or due to lack of signal resulting from fluorescent dye bleaching, or due to the polymerase acting too fast to be detected by device. All of the above events result in a deletion error in the raw sequence. Similarly, mutation errors and insertion errors can also happen at a higher frequency for the simple reasons of potentially weaker signals and faster reactions than in conventional methods.

Low accuracy sequence data is more difficult to assemble. In large scale sequencing, such as sequencing a complete eukaryotic genome, the DNA molecules are fragmented into smaller pieces. These pieces are sequenced in parallel, and then the resultant reads are assembled to reconstruct the whole sequence of the original sample DNA molecules. The fragmentation can be achieved, for example, by mechanical shearing or enzymatic cleavage.

Assembly of small reads of sequence into a large genome requires that the fragmented reads are accurate enough to be correctly grouped together. This is generally true for the raw sequencing data generated from the Sanger method, which can have a raw data accuracy of higher than 95%. Accurate single molecular sequencing technology could be applied to detect single-base modifications or mutations nucleic acid samples. However, the raw data accuracy for single molecule sequencing technologies may be lower due to the limitations discussed above. The accuracy of individual reads of raw sequence data could be as low as 60 to 80%. See, e.g., Harris et al., Science 320:106-109 (2008). Thus, it would be useful to provide accurate single molecule sequencing methods.

Additionally, DNA methylation plays a critical role in the regulation of gene expression; for example, methylation at promoters often leads to transcriptional silencing. Methylation is also known to be an essential mechanism in genomic imprinting and X-chromosome inactivation. However, progress in deciphering complex whole genome methylation profiles has been limited. Therefore, methods of determining DNA methylation profiles in a high-throughput manner could be useful, more so should the methods also provide for accurate determination of sequence.

SUMMARY OF THE INVENTION

In some embodiments, the invention provides a method of determining the sequence of a nucleic acid sample comprising (a) providing a circular nucleic acid molecule comprising at least one insert-sample unit comprising a nucleic acid insert and the nucleic acid sample, wherein the insert has a known sequence; (b) obtaining sequence data comprising sequence of at least two insert-sample units, wherein a nucleic acid molecule is produced that comprises at least two insert-sample units; (c) calculating scores of the sequences of at least two inserts of the sequence data of step (b) by comparing the sequences to the known sequence of the insert; (d) accepting or rejecting at least two of the repeats of the sequence of the nucleic acid sample of the sequence data of step (b) according to the scores of one or both of the sequences of the inserts immediately upstream and downstream of the repeat of the sequence of the nucleic acid sample; (e) compiling an accepted sequence set comprising at least one repeat of the sequence of the nucleic acid sample accepted in step d; and (f) determining the sequence of the nucleic acid sample using the accepted sequence set.

In some embodiments, the invention provides a system comprising a sequencing apparatus operably linked to a computing apparatus comprising a processor, storage, bus system, and at least one user interface element, the storage being encoded with programming comprising an operating system, user interface software, and instructions that, when executed by the processor, optionally with user input, perform a method comprising: (a) obtaining sequence data from a circular nucleic acid molecule comprising at least one insert-sample unit comprising a nucleic acid insert and a nucleic acid sample, wherein: (i) the insert has a known sequence, (ii) the sequence data comprise sequence of at least two insert-sample units, and (iii) a nucleic acid molecule is produced that comprises at least two insert-sample units; (b) calculating scores of the sequences of at least two inserts of the sequence data of step (a) by comparing the sequences to the known sequence of the insert; (c) accepting or rejecting at least two of the repeats of the sequence of the nucleic acid sample of the sequence data of step (a) according to the scores of one or both of the sequences of the inserts immediately upstream and downstream of the repeat of the sequence of the nucleic

acid sample; (d) compiling an accepted sequence set comprising at least one repeat of the sequence of the nucleic acid sample accepted in step (c); and (e) determining the sequence of the nucleic acid sample using the accepted sequence set, wherein an output of the system is used to produce at least one of (i) a sequence of a nucleic acid sample or (ii) an indication that there is a modified base in at least one position in a nucleic acid sample.

In some embodiments, the invention provides a storage encoded with programming comprising an operating system, user interface software, and instructions that, when executed by the processor on a system comprising a sequencing apparatus operably linked to a computing apparatus comprising a processor, storage, bus system, and at least one user interface element, optionally with user input, perform a method comprising: (a) obtaining sequence data from a circular nucleic acid molecule comprising at least one insert-sample unit comprising a nucleic acid insert and a nucleic acid sample, wherein: (i) the insert has a known sequence, (ii) the sequence data comprise sequence of at least two insert-sample units, and (iii) a nucleic acid molecule is produced that comprises at least two insert-sample units; (b) calculating scores of the sequences of at least two inserts of the sequence data of step (a) by comparing the sequences to the known sequence of the insert; (c) accepting or rejecting at least two of the repeats of the sequence of the nucleic acid sample of the sequence data of step (a) according to the scores of one or both of the sequences of the inserts immediately upstream and downstream of the repeat of the sequence of the nucleic acid sample; (d) compiling an accepted sequence set comprising at least one repeat of the sequence of the nucleic acid sample accepted in step (c); and (e) determining the sequence of the nucleic acid sample using the accepted sequence set, wherein the method results in output used to produce at least one of (i) a sequence of a nucleic acid sample or (ii) an indication that there is a modified base in at least one position in a nucleic acid sample.

In some embodiments, the invention provides a method of determining a sequence of a double-stranded nucleic acid sample and a position of at least one modified base in the sequence, comprising: (a) locking the forward and reverse strands together to form a circular pair-locked molecule; (b) obtaining sequence data of the circular pair-locked molecule via single molecule sequencing, wherein the sequence data comprises sequences of the forward and reverse strands of the circular pair-locked molecule; (c) determining the sequence of the double-stranded nucleic acid sample by comparing the sequences of the forward and reverse strands of the circular pair-locked molecule; (d) altering the base-pairing specificity of bases of a specific type in the circular pair-locked molecule to produce an altered circular pair-locked molecule; (e) obtaining the sequence data of the altered circular pair-locked molecule wherein the sequence data comprises sequences of the altered forward and reverse strands; and (f) determining the positions of modified bases in the sequence of the double-stranded nucleic acid sample by comparing the sequences of the altered forward and reverse strands.

In some embodiments, the invention provides a method of determining a sequence of a double-stranded nucleic acid sample, comprising: (a) locking the forward and reverse strands of the nucleic acid sample together to form a circular pair-locked molecule; (b) obtaining sequence data of the circular pair-locked molecule via single molecule sequencing, wherein sequence data comprises sequences of the forward and reverse strands of the circular pair-locked molecule; and (c) determining the sequence of the double-stranded nucleic acid sample by comparing the sequences of the forward and reverse strands of the circular pair-locked molecule.

In some embodiments, the invention provides a method of determining a sequence of a double-stranded nucleic acid sample and a position of at least one modified base in the sequence, comprising: (a) locking the forward and reverse strands of the nucleic acid sample together to form a circular pair-locked molecule; (b) obtaining sequence data of the circular pair-locked molecule via single molecule sequencing, wherein sequence data comprises sequences of the forward and reverse strands of the circular pair-locked molecule; and (c) determining the sequence of the double stranded nucleic acid sample and the position of the at least one modified base in the sequence of the double stranded nucleic acid sample by comparing the sequences of the forward and reverse strands of the circular pair-locked molecule.

In some embodiments, the invention provides a method of determining a sequence of a double-stranded nucleic acid sample and a position of at least one modified base in the sequence, comprising: (a) locking the forward and reverse strands of the nucleic acid sample together to form a circular pair-locked molecule; (b) altering the base-pairing specificity of bases of a specific type in the circular pair-locked molecule; (c) obtaining sequence data of the circular pair-locked molecule via single molecule sequencing, wherein sequence data comprises sequences of the forward and reverse strands of the circular pair-locked molecule; and (d) determining the sequence of the double-stranded nucleic acid sample and the position of the at least one modified base in the sequence of the double-stranded nucleic acid sample by comparing the sequences of the forward and reverse strands of the circular pair-locked molecule.

In some embodiments, the invention provides a method of determining a sequence of a double-stranded nucleic acid sample and a position of at least one modified base in the sequence, comprising: (a) locking the forward and reverse strands together to form a circular pair-locked molecule; (b) obtaining sequence data of the circular pair-locked molecule via single molecule sequencing, wherein the sequence data comprises sequences of the forward and reverse strands of the circular pair-locked molecule; (c) determining the sequence of the double-stranded nucleic acid sample by comparing the sequences of the forward and reverse strands of the circular pair-locked molecule; (d) obtaining sequencing data of the circular pair-locked molecule via single molecule sequencing, wherein at least one nucleotide analog that discriminates between a base and its modified form is used to obtain sequence data comprising at least one position wherein the at least one differentially labeled nucleotide analog was incorporated; and (e) determining the positions of modified bases in the sequence of the double-stranded nucleic acid sample by comparing the sequences of the forward and reverse strands.

In some embodiments, the invention provides a method of determining a sequence of a double-stranded nucleic acid sample and a position of at least one modified base in the sequence, comprising: (a) locking the forward and reverse strands of the nucleic acid sample together to form a circular pair-locked molecule; (b) obtaining sequence data of the circular pair-locked molecule via single molecule sequencing, wherein at least one nucleotide analog that discriminates between a base and its modified form is used to obtain sequence data comprising at least one position wherein the at least one differentially labeled nucleotide analog was incorporated; and (c) determining the sequence of the double-stranded nucleic acid sample and the position of the at least one modified base in the sequence of the double-stranded

nucleic acid sample by comparing the sequences of the forward and reverse strands of the circular pair-locked molecule.

Additional objects and advantages of the invention will be set forth in part in the description which follows, and in part will be obvious from the description, or may be learned by practice of the invention. The objects and advantages of the invention will be realized and attained by means of the elements and combinations particularly pointed out in the appended claims.

It is to be understood that both the foregoing general description and the following detailed description are exemplary and explanatory only and are not restrictive of the invention, as claimed.

The accompanying drawings, which are incorporated in and constitute a part of this specification, illustrate several embodiments of the invention and together with the description, serve to explain the principles of the invention.

BRIEF DESCRIPTION OF THE DRAWINGS

The foregoing aspects and advantages of this invention may become apparent from the following detailed description with reference to the accompanying drawings in which:

FIG. **1**. Preparation of a circular DNA molecule in accordance with some embodiments of the invention. A DNA sample **1** is fragmented; a fragment **2** is ligated at its 5' end (diamond) to a linker **3** and at its 3' end (arrowhead) to another linker **4**. The linkers **3** and **4** are complementary to adjoining segments of an oligonucleotide **5**. Annealing of **5** to **3** and **4** provides a substrate for circularization by ligation, which reaction results in a circular molecule **6** comprising a nucleic acid insert (from the sequence of the linkers **3** and **4**) and a nucleic acid sample (from the sequence of the fragment **2**).

FIG. **2**. Rolling circle amplification. An oligonucleotide **5**, annealed to a circular molecule **6** produced as in FIG. **1**, is bound by a polymerase **7** anchored to a surface **8**. Extension of the oligonucleotide gives a complementary linear copy **9** of the circular molecule. Continued extension results in strand displacement and synthesis of a molecule **10** containing multiple copies of the circular molecule.

FIG. **3**. Circular pair-locked molecule. (A) A double stranded molecule containing a forward strand **11** and reverse strand **12** can be combined with inserts that form hairpins **13** and **14**, which may be identical or non-identical, to form a circular pair locked molecule. In some embodiments, the linkers have overhangs and recessed ends (**37** and **38**). These can be filled in using a polymerase or may be complementary to overhangs in the double stranded molecule (not shown). In a complete circular pair-locked molecule, **37** and **38** are filled in and sealed so that the molecule has a continuous, single stranded, and circular backbone. (B) After gap filling and end joining as appropriate, a circular DNA is formed containing the forward strand **11**, linker **14**, reverse strand **12**, and linker **13**, shown here in melted form. The molecule can be converted to double stranded form, for example, by annealing a primer to one of the linkers and extending it using a polymerase without strand displacement activity, for example, *E. coli* DNA polymerase I, followed by ligation.

FIG. **4**. Schemes for sequence determination and sequence and methylation profile determination using circular pair-locked molecules. (Left) A circular pair locked molecule can be sequenced for at least one full length of the molecule to provide complementary sequence reads; continued sequencing can be used to provide additional redundancy. The sequence data can be aligned and evaluated based on the sequences of the insert nucleic acids so as to obtain accurate sequence of the sample nucleic acid. (Right) Conversion of a specific type of nucleotide, such as by bisulfite conversion or photochemical transition, followed by sequencing, alignment, and comparison of the modified sequence and its unmodified complement can be used to obtain accurate sequence data and methylation profiles. Extended sequence reads containing multiple repeats of the sample nucleic acid sequence can be used for increased accuracy.

FIG. **5**. Nucleotide conversion. (A) A circular pair-locked molecule containing inserts **13** and **14**, a forward strand **15** containing at least one 5-methylcytosine (^{m}C) residue, and a reverse strand **16** is subjected to treatment, such as photochemical transition, to convert ^{m}C to T, resulting in converted forward strand **17**. The complementary nucleotide in the reverse strand is unaffected, resulting in a G-T wobble pair. (^{m}C residues in the reverse strand, if present, would be converted by the treatment.) (B) A circular pair-locked molecule containing inserts **13** and **14**, a forward strand **15** containing at least one 5-methylcytosine (^{m}C) residue, and a reverse strand **16** is subjected to treatment, such as bisulfite conversion, to convert C (but not ^{m}C) to U, resulting in converted forward strand **39** and converted reverse strand **40**. The nucleotides complementary to the converted nucleotides are unaffected, resulting in G-U wobble pairs.

FIG. **6**. Obtaining sequence data and a methylation profile from a circular pair-locked molecule. (A) A primer **18** is annealed to the converted circular pair-locked molecule of FIG. **5**A and extended by a polymerase, resulting in synthesis of a strand with segments **19**, **20**, and **21**, complementary to the sequences of **16**, **14**, and **17**, respectively. (B) Sequence is obtained comprising at least two repeats: at least one of a repeat of the sample **17** and a repeat of the newly synthesized complement of the forward strand **21**; and at least one of a repeat of the newly synthesized complement of the reverse strand **19** and a repeat of the reverse strand **16**. These repeats are aligned; a position **41** at which there is disagreement among the repeats signifies that a base was modified at that position. Depending on the type of modification used, the bases originally present at the corresponding position of the nucleic acid sample can be determined. In this example, where the circular pair locked molecule has been modified by conversion of ^{m}C to T (see FIG. **5**A), the disagreement indicates that a ^{m}C was present in the nucleic acid sample in the forward strand at position **41**; the logic is that at a position where the sequences disagree, the base which is the product of the conversion reaction, T, has replaced the substrate of the conversion reaction, ^{m}C, which was present in the nucleic acid sample.

FIG. **7**. Raw and processed sequence data acquired from a circular nucleic acid molecule template. (A) The content of sequence that can be obtained from a circular template is represented diagrammatically. Nucleic acid sample sequence is represented by dashes and nucleic acid insert sequence is represented by circles. The sequence illustrated begins with a partial sequence **22** of a nucleic acid sample, followed by the sequence of a nucleic acid insert **23**; these are followed by a sequence **24** of the nucleic acid sample, a sequence **25** of a nucleic acid insert, a sequence **26** of the nucleic acid sample, and a sequence **27** of a nucleic acid insert. **28** represents additional sequence not shown in this figure, which is followed by a sequence **29** of a nucleic acid insert, a sequence **30** of the nucleic acid sample, a sequence **31** of a nucleic acid insert, and a partial sequence **32** of a nucleic acid sample.

If the circular template comprises a single nucleic acid sample and a single nucleic acid insert, then both of **22** and **24**, along with subsequent nucleic acid sample sequences **26**, **30**, and **32**, are sequences of the same single nucleic acid sample; likewise, **23**, **25**, **27**, **29**, and **31** are sequences of the same

single nucleic acid insert in this case. If the circular template comprises forward and reverse repeats of the sequence of the nucleic acid sample and two nucleic acid inserts having known sequences, which may be identical or non-identical, as in the case of a circular pair-locked molecule, then the nucleic acid sample sequences have alternating orientations and correspond to the two nucleic acid sample repeats in an alternating manner (e.g., **22** could be in forward orientation, meaning it is a sequence of the reverse repeat, and **24** could be in reverse orientation, meaning it is a sequence of the forward repeat, or vice versa). Likewise, the nucleic acid insert sequences **23**, **25**, etc., would also correspond to the two nucleic acid inserts, which may be identical or non-identical, of the circular template in an alternating manner.

(B) The sequence shown in FIG. **7**A can be decomposed into segments each containing a repeat of the nucleic acid sample sequence, e.g., **24**; the segments also comprise at least one repeat of the nucleic acid insert, for example, two repeats of the nucleic acid insert, e.g., **23** and **25**. Some segments may contain only a partial sequence, e.g., **33**, or an unusually long sequence, e.g., **34**. Such segments can result from errors during sequencing. In some embodiments, such segments are excluded from further consideration.

FIG. **8**. Diagram of sequence processing steps. In some embodiments, raw sequence data are examined, processed, and accepted or rejected as shown. A raw sequence database **35** may be used. If a score is calculated that exceeds a threshold, a step **36** may be performed: accept the sample sequence and add it to an accepted sequence set.

FIG. **9**. Rolling circle amplification products. Products of the reactions described in Example 1 were electrophoresed and the gel was visualized as described. From the left, C1 and C2 are negative control lanes. The leftmost Mr lane contains the FERMENTAS GENERULER 1 kb ladder, Cat. No. SM0311, which band sizes ranging from 250 to 10,000 bp. The next ten lanes contain products of rolling circle amplification reactions as indicated, generated using two primers or one primer (amplification control) and products of the L0 (negative ligation control) or L3 reactions ligation reactions taken at the indicated times; see Example 1. The next Mr lane contains the FERMENTAS GENERULER 100 bp Plus ladder, Cat. No. SM0321, with band sizes ranging from 100 to 3,000 bp. The next ten lanes contain the same products as in the previous ten product lanes except that these products were mixed with loading dye containing 1% SDS.

FIG. **10**. Alignments showing repeat sequences and deduced original sequence of a simulated nucleic acid sample. Positions where all aligned sequences agree are marked by asterisks. (A) Reads a (residues 1 to 35 of SEQ ID NO: 10) and b (residues 1 to 35 of SEQ ID NO: 11) of Example 2 are shown together with the deduced original sequence, labeled 'o' (SEQ ID NO: 5), of the forward strand of the nucleic acid sample. The original sequence was deduced using the rules shown in Table 5. The positions where all three sequences shown have C are positions where the simulated nucleic acid sample contained a methylated cytosine in the forward strand. The positions where all three sequences shown have G are positions where the simulated nucleic acid sample contained a methylated cytosine in the reverse strand. (B) Reads a (SEQ ID NO: 14) and b (SEQ ID NO: 15) of Example 3 are shown together with the deduced original sequence of the forward strand, marked 'r_a' (SEQ ID NO: 5). The original sequence was deduced using the rules in Table 6. The positions where the deduced original sequence has a C that disagrees with read a are positions where the simulated nucleic acid sample contained a methylated cytosine in the forward strand. The positions where the deduced original sequence has a G that disagrees with read b are positions where the simulated nucleic acid sample contained a methylated cytosine in the reverse strand.

FIG. **11**. Computing apparatus and storage. (A) In some embodiments, the invention relates to a sequencing apparatus **51** operably linked to a computing apparatus **52** comprising at least one user interface element chosen from a display **57**, a keyboard **58**, and a mouse **59**, and at least one computer **53** comprising a storage **54** (see panel B), a bus system **55**, and a processor **56**. (B) In some embodiments, the invention relates to a storage **54** comprising an operating system **60**, user interface software **61**, and processing software **62**. The storage can additionally comprise sequence data **63** acquired from the sequencing apparatus (**51** in FIG. **11**A).

FIG. **12**. General scheme of sequence and 5-methylcytosine position determination of using bisulfite conversion with a linear pair locked molecule. A double stranded nucleic acid sample comprising 5-methylcytosine is provided (at top). A linear pair-locked molecule is constructed by ligating a hairpin insert to one double strand end of the molecule (beneath first arrow, at right), thereby locking the forward and reverse strands of the double-stranded sample together. Also, linear flaps are attached to the other double strand end (at left). Bisulfite conversion is performed, converting cytosines to uracils but leaving 5-methylcytosines unaffected. The molecule is copied by providing a primer that binds to the linear flap attached at the 3' end of the linear pair locked molecule and extending the primer with a polymerase. The ends can be processed, e.g., by restriction digestion, to prepare the molecule for subsequent cloning and/or sequencing.

FIG. **13**. General scheme of sequence and 5-methylcytosine position determination using photochemical transition with a linear pair locked molecule. A double stranded nucleic acid sample comprising 5-methylcytosine is provided (at top). A linear pair-locked molecule is constructed by ligating a hairpin insert to one double strand end of the molecule (beneath first arrow, at right), thereby locking the forward and reverse strands of the double-stranded sample together. Also, linear flaps are attached to the other double strand end (at left). Photochemical transition is performed, converting 5-methylcytosines to thymines but leaving unmodified cytosines unaffected. The molecule is copied by providing a primer that binds to the linear flap attached at the 3' end of the linear pair locked molecule and extending the primer with a polymerase. The ends can be processed, e.g., by restriction digestion, to prepare the molecule for subsequent cloning and/or sequencing.

FIG. **14**. General scheme of sequence determination using a circular pair locked molecule. A double stranded nucleic acid sample is provided (at top) top strand: SEQ ID NO: 5; bottom strand: SEQ ID NO: 6). A circular pair-locked molecule is constructed by ligating a hairpin insert to both double strand ends of the molecule (beneath first arrow, at right and left), thereby locking the forward and reverse strands of the double-stranded sample together. Sequencing is performed, resulting in reads of SEQ ID NOs: 5 and 6, and the sequence data is analyzed to determine the sequence of the sample (SEQ ID NOs: 5 and 6); see, e.g., Example 5.

FIG. **15**. General scheme of sequence and 5-methylcytosine position determination using bisulfite conversion and a circular pair locked molecule. A double stranded nucleic acid sample comprising 5-methylcytosine is provided (at top) (top strand: SEQ ID NO: 5; bottom strand: SEQ ID NO: 6). A circular pair-locked molecule is constructed by ligating a hairpin insert to both double strand ends of the molecule (beneath first arrow, at right and left), thereby locking the forward and reverse strands of the double-stranded sample

together. Bisulfite conversion is performed, converting cytosines to uracils but leaving 5-methylcytosines unaffected. The product contains residues 1-35 and 40-74 of SEQ ID NO: 8. Sequencing is performed, resulting in reads of residues 1-35 and 40-74 of SEQ ID NO: 9, and the sequence data is analyzed to determine the sequence of the sample and 5-methylcytosine positions (SEQ ID NOs: 5 and 6); see, e.g., Example 6.

FIG. **16**. General scheme of sequence and 5-methylcytosine position determination using photochemical transition and a circular pair locked molecule. A double stranded nucleic acid sample comprising 5-methylcytosine is provided (at top) (top strand: SEQ ID NO: 5; bottom strand: SEQ ID NO: 6). A circular pair-locked molecule is constructed by ligating a hairpin insert to both double strand ends of the molecule (beneath first arrow, at right and left), thereby locking the forward and reverse strands of the double-stranded sample together. Photochemical transition is performed, converting 5-methylcytosines to thymines but leaving unmodified cytosines unaffected. The product contains residues 1-35 and 40-74 of SEQ ID NO: 12. Sequencing is performed, resulting in reads of residues 1-35 and 40-74 of SEQ ID NO: 13, and the sequence data is analyzed to determine the sequence of the sample and 5-methylcytosine positions (SEQ ID NOs: 5 and 6); see, e.g., Example 7.

FIG. **17**. General scheme of sequence and 5-bromouracil position determination using a circular pair locked molecule. A double stranded nucleic acid sample comprising 5-bromouracil is provided (at top) (top strand: SEQ ID NO: 16; bottom strand: SEQ ID NO: 17). A circular pair-locked molecule is constructed by ligating a hairpin insert to both double strand ends of the molecule (beneath first arrow, at right and left), thereby locking the forward and reverse strands of the double-stranded sample together. Sequencing is performed (resulting in reads of SEQ ID NOs: 18 and 19 and the sequence data is analyzed to determine the sequence of the sample and 5-bromouracil positions (SEQ ID NOs: 16 and 17); see, e.g., Example 8.

DETAILED DESCRIPTION OF THE INVENTION

Definitions

To facilitate the understanding of this invention, a number of terms are defined below. Terms not defined herein have meanings as commonly understood by a person of ordinary skill in the areas relevant to the present invention. Terms such as "a", "an" and "the" are not intended to refer to only a singular entity, but include the general class of which a specific example may be used for illustration. The terminology herein is used to describe specific embodiments of the invention, but their usage does not delimit the invention, except as outlined in the claims.

The term nucleic acid includes oligonucleotides and polynucleotides.

High stringency conditions for hybridization refer to conditions under which two nucleic acids must possess a high degree of homology to each other to hybridize. Examples of high stringency conditions for hybridization include hybridization in 4×sodium chloride/sodium citrate (SSC), at 65 or 70° C., or hybridization in 4×SSC plus 50% formamide at about 42 or 50° C., followed by at least one, at least two, or at least three washes in 1×SSC, at 65 or 70° C.

Melting temperature refers to the temperature at which half of a nucleic acid in solution exists in a melted state and half exists in an unmelted state, assuming the presence of sufficient complementary nucleic acid. In the case of an oligonucleotide present in excess over complementary sequence, melting temperature is the temperature at which half of the complementary sequence is annealed with the oligonucleotide. In the case of a nucleic acid insert capable of forming a hairpin, melting temperature is the temperature at which half of the insert is in a partially self-hybridized "hairpin" form. As melting temperature is condition dependent, melting temperatures of oligonucleotides discussed herein refer to the melting temperature in an aqueous solution of 50 mM sodium chloride, with the oligonucleotide at 0.5 μM. Melting temperatures can be estimated by various methods known in the art, for example, using the nearest-neighbor thermodynamic parameters found in Allawi et al., *Biochemistry,* 36, 10581-10594 (1997) together with standard thermodynamic equations.

A site in a nucleic acid molecule is suitable for primer binding if it has a unique sequence in the nucleic acid molecule and is of a length and composition such that the complementary oligonucleotide has an acceptable melting temperature, for example, a melting temperature ranging from 45° C. to 70° C., from 50° C. to 70° C., from 45° C. to 65° C., from 50° C. to 65° C., from 55° C. to 70° C., from 60° C. to 70° C., from 55° C. to 60° C., from 60° C. to 65° C., or from 50° C. to 55° C.

Extending a primer, oligonucleotide, or nucleic acid refers to adding at least one nucleotide to the primer, oligonucleotide, or nucleic acid. This includes reactions catalyzed by polymerase or ligase activity.

A sequencing primer is an oligonucleotide that can bind to a site in a nucleic acid molecule that is suitable for primer binding and be extended in a sequencing reaction so as to produce sequence data.

A nucleic acid insert is capable of forming a hairpin if it can partially self-hybridize, and the self-hybridized form has a melting temperature of at least 15° C.

An overhang is a single stranded segment at the end of a double stranded nucleic acid molecule or hairpin.

A repeat or repeat sequence is a sequence that occurs more than once in a nucleic acid. When repeats are present in a nucleic acid molecule, all instances of the sequence, including the first instance, are considered repeats. Repeats include sequences that are reverse complements of each other, such as occur in a circular pair-locked molecule. Repeats also include sequences that are not exactly identical but are derived from the same sequence, e.g., sequences that differ due to misincorporation events or other polymerase errors during synthesis, or sequences that were initially identical or perfect reverse complements but differ due to modification by a procedure such as photochemical transition or bisulfite treatment.

A nucleic acid insert and a nucleic acid sample are immediately upstream or downstream of one another if there are no other intervening repeats of the insert or sample between the insert and sample. In a single stranded molecule, upstream refers to the 5' direction and downstream refers to the 3' direction. In a double stranded molecule, the polarity can be determined arbitrarily or it can be determined according to the polarity of directional elements such as promoters, coding sequences, etc., if a majority of such elements is oriented in the same way. The polarity of a promoter is that the direction of an initiating RNA polymerase's synthesis is downstream. The polarity of a coding sequence is that the direction from start to stop codon is downstream.

Two repeats are in forward and reverse orientations relative to each other, and have opposite orientations, if they are reverse complements of each other or one or both are derivatives of the reverse complement of each other. Which repeat is

considered forward can be arbitrary or can be determined according to polarity of elements in the repeat, as discussed in the preceding paragraph.

A modified base is a base other than adenine, thymine, guanine, cytosine, or uracil that can be included in place of one or more of the aforementioned bases in a nucleic acid or nucleotide.

Ambiguity codes are codes that represent a combination of bases at a sequence, in the sense that any of the represented bases could be present, for example: Y=pyrimidine (C, U, or T); R=purine (A or G); W=weak (A, T, or U); S=strong (G or C); K=keto (T, U, or G); M=amino (C or A); D=not C (A, G, T, or U); V=not T or U (A, C, or G); H=not G (A, C, T, or U); B=not A (C, G, T, or U).

A position weight matrix is a matrix in which the rows correspond to positions in a nucleic acid sequence and the columns correspond to bases, or vice versa, and each element in the matrix is a weight for a particular base at a particular position. A sequence can be scored against a position weight matrix by summing the weights corresponding to each base of the sequence; for example, if the sequence is ACG, the score would be the sum of the weight for A in the first column of the matrix, the weight for C in the second column, and the weight for G in the third column, assuming columns corresponded to positions. A position weight matrix can be run over a sequence with a length greater than the number of positions in the matrix by iteratively scoring the sequence against the matrix, in which the starting position is incremented by one position in each run. In this way, a position in the sequence that produces a maximum or minimum score against the matrix can be identified.

Storage refers to a repository of digital information accessible by a computer. It includes RAM, ROM, hard drives, non-volatile solid state memory, optical disks, magnetic disks, and equivalents thereof.

A data structure is an object or variable in a storage that contains data. A data structure can contain scalar data (e.g., an individual character, number, or string), an assembly of scalar data (e.g., a matrix or array of scalars), or a recursive assembly (e.g., a list, which can be multidimensional, comprising sub-lists, matrices, arrays, and/or scalars as elements, with the sub-lists able themselves to contain sub-lists, matrices, arrays, and/or scalars as elements).

Nucleic Acid Sample

The methods of the invention comprise determining the sequence of a nucleic acid sample and/or determining the positions of modified bases in a nucleic acid sample. The term "nucleic acid sample" refers to the nucleic acid whose sequence and/or modified base positions are to be determined in the methods of the invention.

The nucleic acid sample can be obtained from a sources including, without limitation, DNA (including without limitation genomic DNA, cDNA, mtDNA, chloroplast DNA, and extrachromosomal or extracellular DNA) or RNA (including without limitation mRNA, primary transcript RNA, tRNA, rRNA, miRNA, siRNA, and snoRNA). The nucleic acid sample can be from an individual, patient, specimen, cell culture, biofilm, organ, tissue, cell, spore, animal, plant, fungus, protist, bacterium, archaeon, virus, or virion. In some embodiments, the nucleic acid sample is obtained as an environmental sample, e.g., from soil or a body of water; the nucleic acid sample may be obtained as an environmental sample without specific knowledge of whether the nucleic acid is of cellular, extracellular, or viral origin. In addition, the nucleic acid can be obtained from a chemical or enzymatic reaction, including reactions in which synthetic, recombinant, or naturally occurring nucleic acid is modified by an enzyme, for example, a methyltransferase.

In some embodiments, the nucleic acid sample is a processed sample from a source such as one of those listed above. For example, the isolated nucleic acid can be fragmented by shearing, such as by sonication or pipetting through a narrow aperture, or enzymatic digestion, such as with an endonuclease, which can be a restriction endonuclease. In some embodiments, the nucleic acid sample has at least one overhang. The isolated nucleic acid may first be cloned and propagated in a host cell and/or vector, e.g., as a bacterial or yeast artificial chromosome, a minichromosome, plasmid, cosmid, extrachromosomal element, or chromosomally integrated construct.

Providing a Circular Nucleic Acid Molecule

In some embodiments, the methods of the invention comprise providing a circular nucleic acid molecule comprising an insert-sample unit comprising a nucleic acid insert and the nucleic acid sample, wherein the insert has a known sequence. The circular nucleic acid molecule can be single or double stranded.

In some embodiments, the circular nucleic acid molecule is provided by isolating it in circular form from its source, if part of its sequence is known and thus can serve as the nucleic acid insert (e.g., a conserved motif within the sequence of a gene contained in the circular molecule may be known, or the molecule may be known to contain a sequence based on its ability to hybridize under high stringency conditions to another nucleic acid of known sequence). In some embodiments, the sequence of the nucleic acid insert is known only inexactly, as would be the case when knowledge of the sequence is derived from stringent hybridization properties. In some embodiments, the sequence of the nucleic acid insert is known exactly, such as would be the case when the circular nucleic acid molecule has a known backbone sequence or has been engineered to contain a known sequence.

In some embodiments, the circular nucleic acid molecule is provided by performing an in vitro reaction or reactions to incorporate the nucleic acid sample into the circular molecule along with a nucleic acid insert. The in vitro reaction or reactions can in some embodiments comprise ligation by a ligase and/or other strand joining reactions such as can be catalyzed by various enzymes, including recombinases and topoisomerases. DNA ligase or RNA ligase may be used to enzymatically join the two ends of a linear template, with or without an adapter molecule or linkers, to form a circle. For example, T4 RNA ligase couples single-stranded DNA or RNA, as described in Tessier et al., *Anal Biochem,* 158: 171-78 (1986). CIRCLIGASE™ (Epicentre, Madison, Wis.) may also be used to catalyze the ligation of a single stranded nucleic acid. Alternatively, a double stranded ligase, such as *E. coil* or T4 DNA ligase, may be used to perform the circularization reaction.

In some embodiments, providing the circular nucleic acid molecule comprises amplifying a nucleic acid template with primers (which may be random primers with 5' flaps of known sequence that can serve as the nucleic acid insert) comprising complementary regions and circularizing the amplified nucleic acid, such as may be catalyzed by a ligase or a recombinase; the amplified nucleic acid may in some embodiments be processed at its ends, e.g., by restriction or phosphorylation, prior to circularization.

In some embodiments, the circular nucleic acid molecule is provided by performing chemical circularization. Chemical methods employ known coupling agents such as BrCN plus imidazole and a divalent metal, N-cyanoimidazole with $ZnCl_2$, 1-(3-dimethylaminopropyl)-3 ethylcarbodiimide

HCl, and other carbodiimides and carbonyl diimidazoles. The ends of a linear template may also be joined by condensing a 5'-phosphate and a 3'-hydroxyl, or a 5'-hydroxyl and a 3'-phosphate.

In some embodiments, the circular nucleic acid molecule is a circular pair-locked molecule (cPLM). This type of molecule is discussed in detail below.

Providing Forward and Reverse Repeats of the Nucleic Acid Sample; Circular Pair-locked Molecules

In some embodiments, the methods of the invention comprise providing forward and reverse repeats of a nucleic acid sample and locking the forward and reverse strands together to form a cPLM. The general structure of a cPLM is shown in FIG. **3**A. A cPLM is a single-stranded circular nucleic acid molecule that comprises forward and reverse repeats of a nucleic acid sample; the repeats are bracketed by nucleic acid inserts, as shown in FIG. **3**A. The nucleic acid inserts can be identical or non-identical. In some embodiments, the inserts have a length of at least 50 nt or at least 100 nt. In some embodiments, the inserts have a length ranging from 50 or 100 nt to 10,000 or 50,000 nt.

The strands of a linear double stranded nucleic acid sample can be locked together to form a cPLM, for example, by ligating nucleic acid inserts that form hairpins to each end of the molecule. In some embodiments, the nucleic acid inserts that form hairpins have melting temperatures of at least 20° C., 25° C., 30° C., 35° C., 40° C., 45° C., 50° C., 55° C., 60° C., 65° C., or 70° C. The ligation can be blunt-end or sticky-end ligation. Hairpin structures have base-paired stem regions and unpaired loop regions. In some embodiments, the insert nucleic acid comprises a loop region of a size of at least 20, 22, 25, 30, or 35 nucleotides. In some embodiments, this loop region is suitable for primer binding. In some embodiments, the loop region binds a primer with a melting temperature of at least 45° C., 50° C., 55° C., 60° C., 65° C., or 70° C.

In some embodiments, the nucleic acid sample comprises different sticky ends, such as could be generated by digestion using restriction enzymes with different restriction sites, and these different sticky ends favor ligation of different nucleic acid inserts. In some embodiments, the double stranded nucleic acid to be converted in this way can be obtained by extending a random primer comprising a 5' flap of known sequence along a template comprising the desired sample sequence.

The strands of a double stranded nucleic acid can also be locked together to form a cPLM by treatment with an enzyme that converts double-strand ends to hairpins, for example, recombinases that form a phosphotyrosine linkage with one strand of a double stranded molecule followed by hairpin formation through nucleophilic attack on the phosphotyrosine linkage by the other strand. Members of the family, such as λ integrase and Flp recombinase, are examples of such recombinases. See, e.g., Chen et al., *Cell* 69, 647-658 (1992); Roth et al., *Proc Natl Acad Sci USA* 90, 10788-10792 (1993). In some embodiments, the nucleic acid sample comprises recognition sequences for the enzyme that converts double-strand ends to hairpins. In some embodiments, recognition sequences for the enzyme that converts double-strand ends to hairpins are attached to the nucleic acid sample, e.g., by ligation.

In some embodiments, the sample nucleic acid is initially obtained in single stranded form and is converted to double stranded form prior to formation of a cPLM. This can be accomplished, for example, by ligating a hairpin with an overhang to the 3' end of the sample nucleic acid, and then extending from the 3' end of the ligated hairpin to synthesize a complementary strand. A second hairpin can then be joined to the molecule to form a cPLM.

Nucleic Acid Insert

The methods of the invention comprise providing and/or using circular nucleic acid molecules, including cPLMs, comprising at least one nucleic acid insert. In some embodiments, the at least one nucleic acid insert has a partially, inexactly, or completely known sequence, as discussed above. In some embodiments, the sequence of the at least one nucleic acid insert is completely known. In some embodiments, the at least one nucleic acid insert comprises a suitable binding site for an oligonucleotide, including a sequencing primer. In some embodiments, the at least one insert nucleic acid forms a hairpin.

In some embodiments, the at least one nucleic acid insert has a length ranging from 10 to 300, 15 to 250, 30 to 200, or 30 to 100 nucleotide residues. In some embodiments, the at least one nucleic acid insert has a melting temperature ranging from 45° C. to 70° C. or from 50° C. to 65° C.

In some embodiments, the at least one nucleic acid insert comprises a promoter, for example, the T7 RNA polymerase promoter. See, e.g., Guo et al., *J Biol Chem* 280, 14956-14961 (2005). A promoter is recognized by an RNA polymerase as a site for initiating RNA synthesis. Additional promoters are also known in the art.

Insert-sample Unit

The circular nucleic acid molecules used in the methods of the invention comprise at least one nucleic acid sample and at least one nucleic acid insert grouped as at least one insert-sample unit. An insert-sample unit is a segment of nucleic acid in which a nucleic acid insert is immediately upstream or downstream of a nucleic acid sample.

In some embodiments, the circular nucleic acid molecule is a cPLM, which comprises two insert-sample units; the nucleic acid samples in these two units are in opposite orientations to each other, that is, one is a forward repeat of the nucleic acid sample and the other is a reverse repeat. It should be noted that the cPLM may be considered to comprise two insert-sample units wherein the inserts are either upstream or downstream of the samples; that is, a cPLM conforming to the structure shown in FIG. **3**B contains, in order, elements **11** (forward repeat), **14** (insert), **12** (reverse repeat), and **13** (insert), with **13** connecting back to **11** to close the circle. No matter whether the insert-sample units are considered to be **11** with **14**, and **12** with **13**, or **13** with **11**, and **14** with **12**, the molecule contains two insert-sample units. In embodiments in which the orientation of the insert and/or its positioning relative to the sample is functionally significant, e.g., the insert comprises a promoter or a primer binding site, it may be most efficient to group the insert-sample units so as to group the insert with the sample toward which the primer binding site or promoter is oriented, i.e., the sample which would be copied first by a polymerase initiating from the primer binding site or promoter.

Obtaining Sequence Data

Sequencing Method

The methods of the invention comprise obtaining sequence data. In some embodiments, a nucleic acid molecule is produced that comprises at least two insert sample units during the step of obtaining sequence data. In some embodiments, the nucleic acid molecule comprising at least two insert sample units can be produced by synthesizing it from the provided circular nucleic acid molecule. In some embodiments, the nucleic acid molecule comprising at least two insert sample units can be produced by altering the provided circular nucleic acid molecule, e.g., by converting the circular nucleic acid molecule to a linear nucleic acid molecule, which

may be single-stranded in some embodiments. In some embodiments, at least one phosphodiester bond in a nucleic acid molecule, which may be the provided circular nucleic acid molecule or a template synthesis product thereof, is formed or broken in the step of obtaining sequence data.

In some embodiments, sequence data is obtained using a sequencing by synthesis method. In some embodiments, sequence data is obtained using a single molecule sequencing method. In some embodiments, the single molecule sequencing method is chosen from pyrosequencing, reversible terminator sequencing, ligation sequencing, nanopore sequencing, and third-generation sequencing.

In some embodiments, sequence data is obtained using a bulk sequencing method, for example, Sanger sequencing or Maxam-Gilbert sequencing.

Single molecule sequencing methods are distinguished from bulk sequencing methods according to whether a single nucleic acid molecule is isolated as part of the sequencing procedure. The nucleic acid molecule may be single- or double-stranded; two annealed nucleic acid strands are considered a single molecule for this purpose. The isolation of the single molecule may occur in a microwell, via use of a nanopore, by direct or indirect attachment in an optically resolvable manner to a substrate such as a microscope slide, or in any other way that allows sequence data to be obtained from the individual molecule. In indirect attachment, the single molecule is attached to the substrate via a linking structure that binds to the single molecule, for example, a protein or oligonucleotide. Notably, methods in which a single molecule is isolated, then amplified, and sequence data is obtained directly from the amplification product(s) are still considered single molecule methods because a single molecule was isolated and served as the ultimate source of the sequence data. (In contrast, in bulk sequencing methods, a nucleic acid sample is used that contains multiple molecules and data is obtained containing signal that originated from multiple molecules.) In some embodiments, single molecule sequencing is performed wherein redundant sequence is obtained from the same molecule. The redundant sequence can be obtained by sequencing at least two direct or inverted repeats within a molecule, or by sequencing the same segment of the molecule more than once. The redundant sequence can be completely redundant or partially redundant with some variation, e.g., due to differences introduced by alteration of base pairing specificity of bases of a certain type, or due to errors that may occur during the sequencing process. In some embodiments, the alteration of base pairing specificity can occur prior to sequencing. In some embodiments, the same molecule is sequenced multiple times, optionally with an intervening treatment that selectively alters base pairing specificity of bases of a certain type occurring between the iterations of sequencing.

Sanger sequencing, which involves using labeled dideoxy chain terminators, is well known in the art; see, e.g., Sanger et al., *Proc Natl Acad Sci USA* 74, 5463-5467 (1997). Maxam-Gilbert sequencing, which involves performing multiple partial chemical degradation reactions on fractions of the nucleic acid sample followed by detection and analysis of the fragments to infer the sequence, is also well known in the art; see, e.g., Maxam et al., *Proc Natl Acad Sci USA* 74, 560-564 (1977). Another bulk sequencing method is sequencing by hybridization, in which the sequence of a sample is deduced based on its hybridization properties to a plurality of sequences, e.g., on a microarray or gene chip; see, e.g., Drmanac, et al., *Nat Biotechnol* 16, 54-58 (1998).

Single molecule sequencing methods are discussed generally, for example, in Kato, *Int J Clin Exp Med* 2, 193-202 (2009) and references therein.

Pyrosequencing, reversible terminator sequencing, and ligation sequencing are considered to be second-generation sequencing methods. Generally, these methods use amplification products generated from a single molecule, which are spatially segregated from amplification products generated from other molecules. The spatial segregation can be implemented by using an emulsion, a picoliter well, or by attachment to a glass slide. Sequence information is obtained via fluorescence upon incorporation of a nucleotide; after acquiring data, the fluorescence of the newly incorporated nucleotide is eliminated and the process is repeated for the next nucleotide.

In pyrosequencing, the pyrophosphate ion released by the polymerization reaction is reacted with adenosine 5' phosphosulfate by ATP sulfurylase to produce ATP; the ATP then drives the conversion of luciferin to oxyluciferin plus light by luciferase. As the fluorescence is transient, no separate step to eliminate fluorescence is necessary in this method. One type of deoxyribonucleotide triphosphate (dNTP) is added at a time, and sequence information is discerned according to which dNTP generates significant signal at a reaction site. The commercially available Roche GS FLX instrument acquires sequence using this method. This technique and applications thereof are discussed in detail, for example, in Ronaghi et al., *Anal Biochem* 242, 84-89 (1996) and Margulies et al., *Nature* 437, 376-380 (2005) (corrigendum at *Nature* 441, 120 (2006)).

In reversible terminator sequencing, a fluorescent dye-labeled nucleotide analog that is a reversible chain terminator due to the presence of a blocking group is incorporated in a single-base extension reaction. The identity of the base is determined according to the fluorophore; in other words, each base is paired with a different fluorophore. After fluorescence/sequence data is acquired, the fluorophore and the blocking group are chemically removed, and the cycle is repeated to acquire the next base of sequence information. The Illumina GA instrument operates by this method. This technique and applications thereof are discussed in detail, for example, in Ruparel et al., *Proc Natl Acad Sci USA* 102, 5932-5937 (2005), and Harris et al., *Science* 320, 106-109 (2008).

In ligation sequencing, a ligase enzyme is used to join a partially double-stranded oligonucleotide with an overhang to the nucleic acid being sequenced, which has an overhang; in order for ligation to occur, the overhangs must be complementary. The bases in the overhang of the partially double-stranded oligonucleotide can be identified according to a fluorophore conjugated to the partially double-stranded oligonucleotide and/or to a secondary oligonucleotide that hybridizes to another part of the partially double-stranded oligonucleotide. After acquisition of fluorescence data, the ligated complex is cleaved upstream of the ligation site, such as by a type IIs restriction enzyme, for example, Bbvl, which cuts at a site a fixed distance from its recognition site (which was included in the partially double stranded oligonucleotide). This cleavage reaction exposes a new overhang just upstream of the previous overhang, and the process is repeated. This technique and applications thereof are discussed in detail, for example, in Brenner et al., *Nat Biotechnol* 18, 630-634 (2000). In some embodiments, ligation sequencing is adapted to the methods of the invention by obtaining a rolling circle amplification product of a circular nucleic acid molecule, and using the rolling circle amplification product as the template for ligation sequencing.

In nanopore sequencing, a single stranded nucleic acid molecule is threaded through a pore, e.g., using an electrophoretic driving force, and sequence is deduced by analyzing data obtained as the single stranded nucleic acid molecule passes through the pore. The data can be ion current data, wherein each base alters the current, e.g., by partially blocking the current passing through the pore to a different, distinguishable degree.

In third-generation sequencing, a slide with an aluminum coating with many small (~50 nm) holes is used as a zero mode waveguide (see, e.g., Levene et al., *Science* 299, 682-686 (2003)). The aluminum surface is protected from attachment of DNA polymerase by polyphosphonate chemistry, e.g., polyvinylphosphonate chemistry (see, e.g., Korlach et al., *Proc Natl Acad Sci USA* 105, 1176-1181 (2008)). This results in preferential attachment of the DNA polymerase molecules to the exposed silica in the holes of the aluminum coating. This setup allows evanescent wave phenomena to be used to reduce fluorescence background, allowing the use of higher concentrations of fluorescently labeled dNTPs. The fluorophore is attached to the terminal phosphate of the dNTPs, such that fluorescence is released upon incorporation of the dNTP, but the fluorophore does not remain attached to the newly incorporated nucleotide, meaning that the complex is immediately ready for another round of incorporation. By this method, incorporation of dNTPs into an individual primer-template complexes present in the holes of the aluminum coating can be detected. See, e.g., Eid et al., *Science* 323, 133-138 (2009).

Sequencing Template; Amount of Sequencing Data Obtained

In some embodiments, sequence data is obtained directly from a circular nucleic acid molecule, that is, by using the circular nucleic acid molecule as a template. The circular nucleic acid molecule used as a template can be a circular pair-locked molecule. In some embodiments, sequence data is obtained from a product nucleic acid molecule that itself was synthesized using a circular nucleic acid molecule as a template; that is, a template from which sequence data is obtained can be a product nucleic acid molecule synthesized from a circular nucleic acid molecule template. In some embodiments, sequence data is obtained from both a circular nucleic acid molecule template and from a product nucleic acid molecule synthesized from the circular nucleic acid molecule template.

In some embodiments, rolling circle amplification, comprising synthesizing a product nucleic acid molecule comprising at least two insert-sample units using the circular nucleic acid molecule as a template, is performed. In some embodiments, the rolling circle amplification comprises synthesizing a product nucleic acid molecule comprising at least 3, 4, 5, 10, 15, 20, 25, 50, or 100 insert-sample units. The use of rolling circle amplification to produce a number of copies of a template is well known in the art; see, e.g., Blanco et al., *J Biol Chem* 264, 8935-8940 (1989) and Banér et al., *Nucleic Acids Res* 26, 5073-5078 (1998). The rolling circle amplification can be performed as part of sequencing in which the circular nucleic acid molecule is the sequencing template, or to synthesize a product nucleic acid molecule which is to be used as a sequencing template.

Regardless of the template, the sequence data obtained according to the methods of the invention comprises at least two repeats of the nucleic acid sample sequence; these at least two repeats can include, in some embodiments, at least one forward repeat of the nucleic acid sample sequence and at least one reverse repeat of the nucleic acid sample sequence. In some embodiments, the sequence data comprise at least 3, 4, 5, 10, 15, 20, 25, 50, or 100 repeats of the nucleic acid sample sequence. In some embodiments, the sequence data comprise at least 2, 3, 4, 5, 10, 15, 20, 25, 50, or 100 forward repeats of the nucleic acid sample sequence. In some embodiments, the sequence data comprise at least 2, 3, 4, 5, 10, 15, 20, 25, 50, or 100 reverse repeats of the nucleic acid sample sequence. In some embodiments, the sequence data comprise at least 2, 3, 4, 5, 10, 15, 20, 25, 50, or 100 each of forward and reverse repeats of the nucleic acid sample sequence.

Calculating Scores

In some embodiments, the methods of the invention comprise calculating scores of the sequences of at least two inserts in the sequence data by comparing the sequences to the known sequence of the insert. In embodiments in which the sequence of the insert is only partially or inexactly known, the known sequence of the nucleic acid insert can comprise ambiguous or unknown positions, for example, through use of ambiguity codes or a position weight matrix.

Comparing the sequences to the known sequence of the insert includes identifying the sequences of at least two inserts in the sequence data. Identifying the sequences can be done in some embodiments by visual inspection, i.e., by a person visually scanning the sequence data and spotting the insert nucleic acid sequences contained therein, or by a computer-aided alignment method. See, e.g., International Patent Application Publication WO 2009/017678. In some embodiments, identifying the sequences can be done by scanning the sequence data using an algorithm that recognizes the sequences, for example, by calculating scores iteratively or heuristically for multiple positions within the sequence data in order to identify local extrema that correspond most closely to the known sequence of the nucleic acid insert. In some embodiments, identifying the sequence of the at least two nucleic acid inserts is performed simultaneously with calculating the scores, in that both processes can utilize the same score.

In some embodiments, calculating scores comprises performing an alignment using an appropriate alignment algorithm, of which many are known in the art and are readily available, for example, BLAST, MEGABLAST, Smith-Waterman alignment, and Needleman-Wunsch alignment. See, e.g., Altschul et al., *J Mol Biol* 215, 403-410 (1990). Appropriate alignment algorithms include both algorithms allowing gaps and algorithms that do not allow gaps. Alternatively, in some embodiments, calculating scores comprises analyzing the sequences using an algorithm such as running a position weight matrix over the sequences and calculating the sum of the elements of the matrix corresponding to the sequence. In this way, the score can be calculated as the local maximum found by applying the matrix to a sequence read in a stepwise fashion.

In some embodiments, the scores are positively correlated with the closeness of the at least two nucleic acid insert sequences to the known sequence (e.g., the maximum possible score results from an exact match). Such positively correlated scores include, without limitation, percent identity, bit scores, and matching base count.

In some embodiments, the scores are negatively correlated with the closeness of the at least two nucleic acid insert sequences to the known sequence (e.g., the minimum possible score results from an exact match). Such negatively correlated scores include, without limitation, e-value, number of mismatches, number of mismatches and gaps, percent mismatched, and percent mismatched/gapped.

In some embodiments, the scores are calculated on a rate basis. The possible range of scores calculated on a rate basis does not change as a function of the length of the sequences

being compared. Examples of scores calculated on a rate basis include, without limitation, percent identity and percent mismatched/gapped.

In some embodiments, the scores are calculated on a count basis. The possible range of scores calculated on a count basis changes as a function of the length of the sequences being compared. Examples of scores calculated on a count basis include, without limitation, bit scores, number of mismatches, number of mismatches and gaps, and matching base count.

Accepting or Rejecting Repeats of the Sequence of the Nucleic Acid Sample; Accepted Sequence Set

In some embodiments, the methods of the invention comprise accepting or rejecting repeats of the sequence of the nucleic acid sample in the sequence data according to the scores of one or both of the sequences of the inserts immediately upstream and downstream of the repeat of the sequence of the nucleic acid sample. Thus, in various embodiments, the scores of both the immediately upstream and immediately downstream nucleic acid inserts, the score of either one, or the score of one or the other specifically is/are used to decide whether to accept or reject a nucleic acid sample sequence in the sequence data.

In embodiments in which the scores are positively correlated with the closeness of the at least two nucleic acid insert sequences to the known sequence, scores are required to be greater than, or greater than or equal to, a threshold value in order to accept a sequence. The choice of an appropriate threshold value depends on multiple factors, including the type of score being used, the error rate of the sequencing method, and time and redundancy considerations.

Accepting and rejecting repeats of the sequence of the nucleic acid sample can be implemented in various ways such that at least one accepted repeat is used, and any rejected repeats are not used, to determine the sequence of the nucleic acid sample. Accepting and rejecting repeats may or may not be performed in a concerted manner with compiling an accepted sequence set. For example, the sequences of accepted repeats can be copied as they are accepted into a new data structure, which becomes the accepted sequence set. Or, the sequences of rejected repeats can be deleted or overwritten (e.g., with '0' or 'X' characters that represent null or excluded data) as they are rejected; in this case, once the rejected sequences have been deleted or overwritten, the original data structure has been modified so as to become the accepted sequence set. In these examples, accepting and rejecting repeats is considered to be performed in a concerted manner with compiling an accepted sequence set.

In some embodiments, repeats of the sequence of the nucleic acid sample can be rejected on an additional basis, such as having a length that deviates from the length of other repeats of the sequence of the nucleic acid sample (see, e.g., FIG. **7**B). For example, a repeat of the sequence of the nucleic acid sample can be rejected if it deviates to a threshold extent from the mean or median length of the other nucleic acid sample sequences, or of a preliminary version of the accepted sequence set comprising repeats of the sequence of the nucleic acid sample accepted according to the scores of one or both of the sequences of the inserts immediately upstream and downstream of the repeat of the sequence of the nucleic acid sample as described above, which may or may not have the repeat of the sequence of the nucleic acid sample under consideration for possible rejection temporarily removed for calculation of the median or mean length. The threshold extent can be expressed in terms of absolute length, for example, 1, 2, 5, 10, 20, or 50 nucleotides; relative length, for example, 1%, 2%, 5%, 10%, 20%, or 50%; or in terms of a statistical measure, such as standard deviation, for example, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, or 5 standard deviations.

Alternatively, the sequences can be flagged as accepted or rejected, and then after the flagging process is complete, the accepted sequences can be copied into a new data structure, or the rejected sequences can be deleted or overwritten, to generate an accepted sequence set in a non-concerted manner.

The accepted sequence set can be chosen from forms including a single data string, which comprises the at least one accepted repeat of the sequence of the nucleic acid sample and any additional accepted repeats in concatenated form, and a multi-element variable, in which each element represents an accepted repeat of the sequence of the nucleic acid sample or a subpart thereof. In some embodiments, the multi-element variable is chosen from a list, array, hash, and matrix. Any form of data structure allowing for storage of the at least one accepted repeat of the sequence of the nucleic acid sample and subsequent determination of the sequence of the nucleic acid sample is suitable for use.

In embodiments in which the form of the accepted sequence set differs from the form of the raw sequence data (e.g., the raw sequence data is in the form of a string and the accepted sequence set is in the form of a multi-element data structure such as an array), the raw sequence data can be parsed into elements containing repeats, insert-sample units, or sample repeats flanked by the immediately upstream and downstream inserts at a point in the method after the raw sequence data is obtained and before the final accepted sequence set is generated. This parsing step can occur before or after the scoring step discussed above.

Determining the Sequence of the Nucleic Acid Sample; Consensus Sequences; Confidence Levels

In some embodiments, the methods comprise determining the sequence of the nucleic acid sample.

The mode of determining the sequence of the nucleic acid sample can be chosen conditionally based on the number of repeats of the nucleic acid sample in the accepted sequence set. For example, when the accepted sequence set contains only one accepted repeat, the sequence of the nucleic acid sample can be determined to be the sequence of the accepted repeat. When the accepted sequence set contains only two, or at least three, accepted repeats, the sequence of the nucleic acid sample can be determined to be the consensus sequence (see below) of the accepted repeats. More options for how the consensus sequence is determined are available when the accepted sequence set contains at least three accepted repeats.

Consensus Sequence

The consensus sequence is determined from an alignment (performed as discussed above, in the "Calculating scores" section) of the accepted repeats; at positions in the alignment where the accepted repeats contain the same base, the consensus sequence contains that base. In some embodiments, at positions in the alignment where the accepted repeats do not contain the same base, the consensus sequence contains the appropriate ambiguity code (e.g., R when the accepted repeats contain A and G at a position). In some embodiments, at positions in the alignment where the accepted repeats do not contain the same base, the consensus sequence contains an N or other symbol indicative of an unknown base. In some embodiments, at positions in the alignment where the accepted repeats do not contain the same base, the consensus sequence contains the base from the accepted repeat that gave a stronger or more robust signal during acquisition of the sequence (e.g., if the raw data were in the form of fluorescence, the base which was called based on brighter fluores-

cence emission (in some embodiments, after appropriate normalization and/or standardization) is placed in the consensus sequence.

When a consensus sequence is determined from an accepted sequence set containing at least three accepted repeats, the base at each position of the consensus sequence can in some embodiments be determined by majority vote; i.e., the base present at a position in more than half of the accepted repeats is placed at that position in the consensus sequence. When the accepted repeats disagree at a position such that there is no majority vote at that position, the base at that position in the consensus sequence is determined by another method, for example, the plurality vote can be used (i.e., the base most frequently present at a position in of the accepted repeats is placed at that position in the consensus sequence), or one of the procedures discussed in the preceding paragraph can be used.

In some embodiments, when a consensus sequence is determined from an accepted sequence set containing at least three accepted repeats, the base at each position of the consensus sequence can in some embodiments be determined according to the frequency of each base at that position in the accepted repeats. Thus, the consensus sequence can be a probabilistic representation of the likelihood that each base is present at each position in the nucleic acid sample. Such a representation can take the form of a position weight matrix. In some embodiments, the elements of the position weight matrix are the frequencies with which each base was observed at each position in the alignment of the accepted repeats.

In some embodiments, the elements of the position weight matrix are calculated from the frequencies with which each base was observed at each position in the alignment of the accepted repeats; other factors can also be used in this calculation, for example, when some accepted repeat sequences were acquired with stronger or more robust signals during acquisition of the sequence than other repeats, the accepted repeat sequences can be given more weight, and/or the other repeats can be given less weight. The degree to which the weights are modified can be quantitatively determined, based, for example, on the signal strength, or it can be a fixed modification; for example, the weight of bases acquired with a relatively strong signal can be increased by a value such as 50% or 100%, and/or the weight of bases with a relatively weak signal can be reduced by a value such as 33% or 50%.

In some embodiments, the elements of the position weight matrix are values which have been derived from transformed frequencies of each base at each position (possibly weighted as discussed above). Frequencies can be transformed, for example, logarithmically or by exponentiation; in some embodiments, the transformation has the effect of down weighting bases rarely observed at a position and/or up weighting the base or bases commonly observed at a position. For example, if T is present at a position in an alignment of N accepted repeat sequences M times, where $N>2$ and $M<N/2$, and C is present every other time (i.e., N−M times), in some embodiments the transformation of these frequencies would result in the weight of T in the position weight matrix being less than M/N (or the percentage corresponding thereto) and/or the weight of C being greater than (N−M)/N (or the percentage corresponding thereto). In some embodiments, the transformation is chosen so as to only up weight the most commonly observed base (or bases in the case of a tie in frequency).

Confidence Levels

In some embodiments, a confidence level is determined for at least one position in the sequence of the nucleic acid sample. A confidence level can be expressed in a number of ways, for example, as an overall base call accuracy value, expressed as a percentage or as a phred score, or as an error rate. In some embodiments, the confidence level is determined from the frequency of the most common base or bases at a position, or of the combined frequency of the bases that are not the most common. In some embodiments, these frequencies are transformed, up weighted, and/or down weighted as discussed above.

Determining a Confidence Level of the Sequence as a Whole; Determining the Sequence of the Nucleic Acid Sample and Confidence Levels in Real Time and/or to a Desired Level of Confidence

In some embodiments, the methods of the invention comprise determining a confidence level of the sequence as a whole. The confidence level of the sequence as a whole can be expressed in a number of ways, for example, as an overall base call accuracy value, expressed as a percentage or as a phred score; as an error rate; or as an expected number of errors in the sequence.

Confidence levels from the individual positions, as discussed in the above section, can be used to calculate the confidence level of the sequence as a whole. For example, an overall confidence level can be determined as the arithmetic mean, geometric mean, median, or modal confidence level of the statistical population of confidence levels at each position of the sequence of the nucleic acid sample. In some embodiments, the statistical population of confidence levels at each position of the sequence of the nucleic acid sample is processed prior to calculation of the confidence level of the sequence as a whole, for example, to reject outliers.

In some embodiments, the methods of the invention comprise determining the sequence of the nucleic acid sample and confidence levels in real time. In these embodiments, data acquired in the sequencing step is processed to determine sequence and confidence levels concurrently with the acquisition of additional sequence data, e.g., from additional repeats of a rolling circle amplification product. As the additional sequence data is acquired, both the determined sequence and the confidence levels are updated. In some embodiments, the real time process is continued until a preselected confidence level is reached. The preselected confidence level can be, for example, a base call accuracy of 90%, 95%, 99%, 99.5%, 99.9%, 99.95%, or 99.99%. The preselected confidence level can be for the sequence as a whole or a fraction of the positions in the sequence, and can be chosen from values such as, for example, 50%, 67%, 75%, 80%, 85%, 90%, 95%, 98%, 99%, 99.5%, and 99.9%.

Multiple Samples; Assembling a Contig

In some embodiments, the method comprises repeating the steps of the method using at least one other sample from the same source, species, or strain as the nucleic acid sample that has a sequence, that partially overlaps the sequence of the nucleic acid sample, thereby determining at least one other sequence, and assembling the at least one other sequence with the sequence of the original sample to form a contig. In some embodiments, the method comprises repeating the steps of the method with many samples, so as to generate contigs of sizes greater than 0.5, 1, 2, 5, 10, or 100 kb, or 1, 2, 5, 10, 100, or 1,000 Mb. In some embodiments, the contig represents the complete sequence, or the complete sequence except for heterochromatic or refractory regions, of a nucleic acid molecule, which may be, for example and without limitation, a chromosome, minichromosome, artificial chromosome, viral genome, or extrachromosomal element. Contig assembly can be carried out using methods known in the art.

Modified Bases

In some embodiments, the nucleic acid sample comprises at least one modified base, for example, 5-methylcytosine, 5-bromouracil, uracil, 5,6-dihydrouracil, ribothymine, 7-methylguanine, hypoxanthine, or xanthine. Uracil can be con-

sidered a modified base in a DNA strand, and ribothymine can be considered a modified base in an RNA strand. In some embodiments, at least one modified base in the double-stranded nucleic acid sample is paired with a base that has a base pairing specificity different from the preferred partner base(s) of the modified base. This can occur, for example, when one base in a double stranded molecule has undergone a reaction (e.g., due to sporadic oxidation, or exposure to a mutagenizing agent such as radiation or a chemical mutagen) that converted it from one of the standard bases to a modified base that does not have the same preferred partner base(s).

Preferred partner bases are based on Watson-Crick base pairing rules. For example, the preferred partner base of adenine is thymine (or uracil), and vice versa; the preferred partner base of cytosine is guanine, and vice versa. Preferred partner bases of modified bases are generally known to those of skill in the art or can be predicted based on the presence of hydrogen bond donors and acceptors in positions analogous to those of the standard bases. For example, hypoxanthine has a hydrogen bond acceptor (a double-bonded oxygen) in the 6 position of the purine ring, like guanine, and therefore its preferred partner base is cytosine, which has a hydrogen bond donor (an amine group) in the 6 position of the pyrimidine ring. Notably, hypoxanthine can be formed by deamination of adenine. As adenine would normally be paired with thymine in DNA, this deamination reaction can result in a hypoxanthine-thymine pair, in which the modified base hypoxanthine is not paired to its preferred partner base. Cytosine can also be deaminated to form uracil. In the context of DNA, uracil can be considered a modified base, and if it is paired to guanine (as can result from cytosine deamination in normal double-stranded DNA), then this is also a situation where the modified base uracil is not paired to its preferred partner base.

Detection of Modified Bases; Altering the Base Pairing Specificity of Bases of a Specific Type

In some embodiments, the methods of the invention comprise altering the base pairing specificity of bases of a specific type. Altering the base pairing specificity of bases of a specific type can comprise specifically altering the base pairing specificity of an unmodified version of a base, e.g., cytosine. In this case, the base pairing specificity of at least one modified form of the base, for example, 5-methylcytosine, is not altered.

Alternatively, altering the base pairing specificity of bases of a specific type can comprise specifically altering the base pairing specificity of a modified version of a base (e.g., 5-methylcytosine), but not the unmodified version of the base (cytosine).

In some embodiments, altering the base pairing specificity of bases of a specific type comprises photochemical transition, which converts 5-methylcytosine (but not unmodified cytosine) to thymine. See, e.g., Matsumura et al., *Nucleic Acids Symp Ser* No. 51, 233-234 (2007). This reaction alters the base pairing specificity of the bases undergoing photochemical transition from guanine to adenine (guanine pairs with 5-methylcytosine while adenine pairs with thymine).

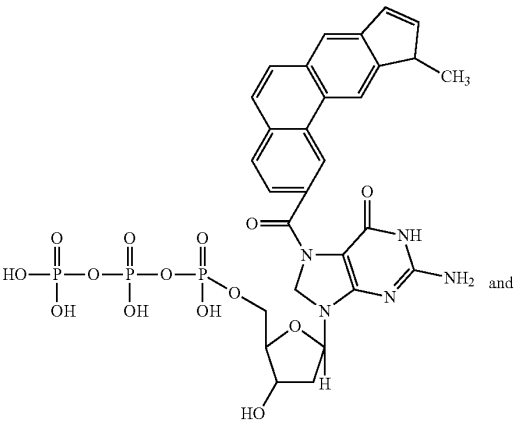
In other embodiments, altering the base pairing specificity of bases of a specific type comprises bisulfite conversion, which converts cytosine (but not 5-methylcytosine) to uracil. See, e.g., Laird et al., *Proc Natl Acad Sci USA* 101, 204-209 (2004), and Zilberman et al., *Development* 134, 3959-3965 (2007). This reaction alters the base pairing specificity of the bases undergoing bisulfite conversion from guanine to adenine (guanine pairs with cytosine while adenine pairs with uracil).

In still other embodiments, modified bases can be detected without an alteration step, such as in cases where the modified base has altered base pairing specificity relative to the unmodified version of the base. Examples of such bases may include 5-bromouracil, uracil, 5,6-dihydrouracil, ribothymine, 7-methylguanine, hypoxanthine, and xanthine. See, e.g., Brown, *Genomes*, 2nd Ed., John Wiley & Sons, Inc., New York, N.Y., 2002, chapter 14, "Mutation, Repair, and Recombination," discussing the propensity of 5-bromouracil to undergo keto-enol tautomerization which results in increased pairing to guanine relative to adenine, and the formation of hypoxanthine (which pairs preferentially to cytosine over thymine) by deamination of adenine.

Nucleotide Analog that Discriminates Between a Base and its Modified Form

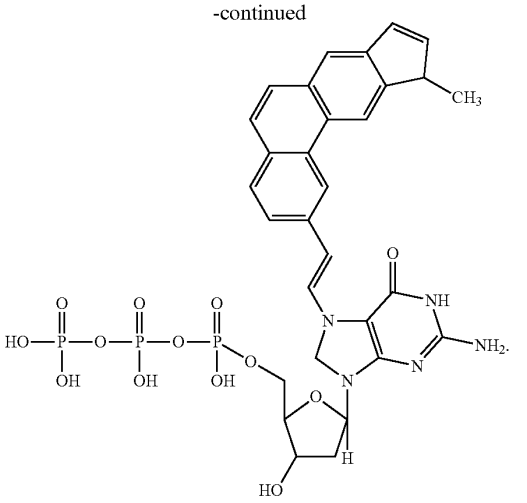
In some embodiments, sequence data is obtained using at least one nucleotide analog that discriminates between a base and its modified form (a "discriminating analog"; it pairs preferentially with one but not the other of the base and its modified form). The nucleotide analog can be used and detected as though it is a fifth base in addition to the standard four bases, for example, by use of differential labels in reversible terminator sequencing or ligation sequencing, or when it is incorporated in pyrosequencing, in which nucleotides can be added one at a time and then washed away. In some embodiments, the discriminating analog is added before its corresponding natural nucleotide (e.g., in pyrosequencing) or provided in a concentration ranging from 10 to 100-fold higher than the concentration of its cognate natural nucleotide (e.g., in reversible terminator sequencing). For example, the discriminating analog can be an analog of deoxyguanosine triphosphate that discriminates between cytosine and 5-methylcytosine (e.g., it will pair with cytosine but not 5-methylcytosine); the analog can be provided at a concentration ranging from 10 to 100-fold higher than the concentration of deoxyguanosine triphosphate. In this way, the analog should generally be incorporated opposite the version of the base it preferentially pairs with, but the natural base should generally be incorporated opposite the version of the base that the analog does not preferentially pair with.

Examples of discriminating analogs can be found in U.S. Pat. No. 7,399,614, and include, for instance, the following molecules, which discriminate between unmodified cytosine and 5-methylcytosine, in that they preferentially pair with the former:



and

-continued



These molecules are referred to as Discriminating Analog 1 and Discriminating Analog 2, respectively.

Determining the Positions of Modified Bases in the Nucleic Acid Sample

In some embodiments, the methods of the invention comprise determining the positions of modified bases in the nucleic acid sample. These embodiments comprise (i) providing the nucleic acid sample in double-stranded form; (ii) converting the nucleic acid sample into a circular pair-locked molecule, wherein the circular pair-locked molecule comprises forward and reverse repeats of the sequence of the nucleic acid sample and two nucleic acid inserts having known sequences, which may be identical or non-identical; (iii) optionally altering the base-pairing specificity of bases of a specific type in the circular pair-locked molecule; (iv) then, obtaining sequence data templated by the forward and reverse repeats of the circular pair-locked molecule or by a complementary sequence thereof; and (v) determining the positions of the modified bases in the nucleic acid sample using the sequence data of at least the forward and reverse repeats or copies thereof. Notably, sequence templated by a forward repeat will have the same sense as the reverse repeat (and vice versa), but may or may not be completely identical to the reverse repeat; differences can result from the forward repeat containing bases that can pair to a base other than the corresponding base in the reverse repeat. An example of such a situation is if the forward repeat in a cPLM contains 5-bromouracil which had been paired to an adenine in the reverse strand but templates the addition of a guanine in a sequencing-by-synthesis reaction.

Sequence data are obtained comprising at least two repeats: at least one of a repeat of the sample (e.g., the repeat labeled **17** in FIG. **5**A) and a repeat of the newly synthesized complement of the forward strand (e.g., the repeat labeled **21** in FIG. **6**A); and at least one of a repeat of the newly synthesized complement of the reverse strand (e.g., the repeat labeled **19** in FIG. **6**A) and a repeat of the reverse strand (e.g., the repeat labeled **16** in FIG. **6**A). These repeats are aligned. The alignment can be performed using any appropriate algorithm, as discussed above. A position at which there is disagreement among the repeats (e.g., the position labeled **41** in FIG. **6**B) signifies that a base in the nucleic acid sample at that position underwent alteration of its base pairing specificity. Depending on the type of modification, modified base, and/or discriminating analog used in the process or present in the sample, the bases originally present at the corresponding position of the nucleic acid sample can be determined.

For example, where the circular pair locked molecule has been altered by conversion of ^{m}C to T (see FIG. **5**A), the disagreement indicates that a ^{m}C was present in the nucleic acid sample at the position that is a T or complementary to an A in one read, and is a C or complementary to a G in another read; the logic is that at a position where the sequences disagree, the base which is the product of the conversion reaction, T, has replaced the substrate of the conversion reaction, ^{m}C, which was present in the nucleic acid sample.

In another example, where the circular pair locked molecule has been altered by conversion of C to U, the disagreement indicates that a C was present in the nucleic acid sample at the position occupied by that is a U or T, or is complementary to an A in one read, and is a C or complementary to a G in another read; the logic is that at a position where the sequences disagree, the base which is the product of the conversion reaction, U (which may be read by the sequencing system as a T), has replaced the substrate of the conversion reaction, C, which was present in the nucleic acid sample. As ^{m}C residues would not be changed by conversion of C to U, the positions where the reads are in agreement in showing C at a position and/or G as its complement indicate that ^{m}C was present at this position in the original sample.

In embodiments in which a discriminating analog was used as discussed above, the presence of the base it preferentially binds to can be inferred in the original sequence at the position of the original sequence corresponding to the position where the discriminating analog appears.

System/Computer Readable Medium

In some embodiments, the invention relates to a system comprising a sequencing apparatus operably linked to a computing apparatus comprising a processor, storage, bus system, and at least one user interface element. The user interface element can be chosen from a display, a keyboard, and a mouse. In some embodiments, the system comprises at least one integrated circuit and/or at least one semiconductor.

In some embodiments, the sequencing apparatus is chosen from sequencing apparatuses configured to perform at least one of the sequencing methods discussed above.

In some embodiments, the display can be a touch screen, serving as the sole user interface element. The storage is encoded with programming comprising an operating system, user interface software, and instructions that, when executed by the processor on a system comprising a sequencing apparatus operably linked to a computing apparatus comprising a processor, storage, bus system, and at least one user interface element, optionally with user input, perform a method of the invention as described above. In some embodiments, the storage further comprises sequence data, which can be in any of the forms discussed above, for example, raw sequence data, an accepted sequence set, a consensus sequence, etc.

In some embodiments, the storage and all of its contents are located within a single computer. In other embodiments, the storage is divided between at least two computers, for example, computers linked via a network connection. In some embodiments, the user interface is part of one computer which is in communication with at least one other computer comprising at least one component of the system, for example, the processing software.

In some embodiments, output of a system or a method executed by a processor results in an indication that there is a modified base in at least one position in a nucleic acid sample. The indication can be in any number of forms, for example, a list of the modified positions in the sequence, a textual or graphical representation of the sequence wherein the modified positions are highlighted or marked, such as with an asterisk or similar character or with bold, italic, or underline formatting, colored text, or a depiction of the chemical structure of the nucleic acid including the structure of the modified base.

EXAMPLES

The following specific examples are to be construed as merely illustrative, and not limitative of the remainder of the disclosure in any way whatsoever. Without further elaboration, it is believed that one skilled in the art can, based on the description herein, utilize the present invention to its fullest extent.

Example 1

Rolling Circle Amplification of a Synthetic Circular Pair-locked Molecule

Four oligodeoxyribonucleotides were provided, as shown in Table 1.

TABLE 1

Oligonucleotide sequences

Name	Sequence	SEQ ID NO
CPLM-1	CGACTTATGCATTTGGTATCTGCGCTCTGCATATTTAAATGGAAGGAGATAGTTAAGGATAAGGGCAGAGCGCAGATAC	1
CPLM-2	CAAATGCATAAGTCGTGTCTTACCGGGTTGATAGCGGCCGCTCGGAGAAAAGAAGTTGGATGATGCAACCCGGTAAGACA	2
pS-T1	CCTTATCCTTAACTATCTCCTT	3
pS-T2	TAGCGGCCGCTCGGAGAAAAG	4

CPLM-1 and CPLM-2 were phosphorylated in separate 50 μL reactions in which 30 μL of 10 μM oligodeoxyribonucleotide (final concentration) was treated with 1 μL of 10 U/μL T4 polynucleotide kinase (New England Biolabs ("NEB") Cat. No. M0201S), in the presence of 5 μL 10× T4 ligase buffer (NEB; the 10× stock buffer contains 10 mM ATP). 14 μL ddH_2O were added to give a final volume of 50 μL (see Table 2). The reactions were incubated at 37° C. for 30 min, followed by enzyme inactivation at 65° C. for 20 min.

TABLE 2

Phosphorylation reaction conditions (volumes in μL)

Reagent	5'P-CPLM-1	5'P-CPLM-2
10 uM CPLM-1	30	0
10 uM CPLM-2	0	30
10 u/uL T4 PNK	1	1
10 × T4 Ligase buffer	5	5
ddH_2O	14	14
Total volume	50	50

The concentration of phosphorylated CPLM-1 and CPLM-2 (5'P-CPLM-1 and 5'P-CPLM-2, respectively) from the above reactions was adjusted to 6 μM.

Phosphorylated CPLM-1 and CPLM-2 were then denatured at 95° C. for 5 min, then placed on ice and mixed with buffer, ddH_2O, and T4 ligase (NEB, Cat. No. M0202S) to produce circular pair-locked molecules, as shown in Table 3. The ligation occurred at 25° C., and 18 μL aliquots were removed at 10, 30 and 60 min. A negative control with no ligase was run in parallel (L0 column in Table 3).

TABLE 3

Ligation reaction conditions

Reagent	L0	L3
6 μM 5'P-CPLM-1	9	9
6 μM 5'P-CPLM-2	9	9
400 u/μL T4 Ligase	0	3
10 × buffer	6	6
ddH_2O	36	33
Total volume	60	60

The ligation products were combined with pS-T1 and/or pS-T2 primers, dNTPs, RepliPHI™ Phi29 DNA polymerase (Epicentre, Cat. No. PP031010), and an appropriate 10× reaction buffer RepliPHI Phi29 DNA Polymerase buffer as shown in Table 4.

TABLE 4

Rolling circle amplification of circular pair-locked molecules

	Controls		2-primed		1-primed	
Reagent	C1	C2	L0	L3	L0	L3
10 mM dNTP	5	5	5	5	5	5
10 μM pS-T1 primer	0	0	6	6	0	0
10 μM pS-T2 primer	0	0	6	6	6	6
1 × L0_10, 30, 60 min	1	0	1	0	1	0
1 × L3_10, 30, 60 min	0	1	0	1	0	1
1000 u/μL phi29 polymerase	1	1	1	1	1	1
10 × buffer	5	5	5	5	5	5
ddH_2O	38	38	26	26	32	32
Total volume	50	50	50	50	50	50

The reactions were assembled without Phi29 polymerase, denatured at 95° C. for 5 min, and placed on ice for 5 min. Phi29 polymerase was added followed by incubation at 30° C. for 18 hours.

5 μL samples of reaction products were mixed with 1 μL 6× loading dye (0.03% bromophenol blue, 0.03% xylene cyanol FF, 60% glycerol, 100 mM Tris-EDTA (pH 7.6)), heated at 95° C. for 10 min, and then placed on ice immediately. A second set of reaction product samples was treated identically except that 1% SDS was added as well.

Samples were loaded into a 0.7% agarose gel in 1× TAE buffer and electrophoresed at 135 V for 28 min. DNA was visualized using GelRed™ precast gel staining (Biotium, Cat. No.: 41003 GelRed™ Nucleic Acid Gel Stain, diluted 10,000× in water). The gel is shown in FIG. **9**. Rolling circle amplification products with apparent molecular weights greater than 10 kb were observed in the samples from reactions using L3 ligation reaction products and both pS-T1 and pS-T2 primers, but not the samples using the L0 controls or the samples that lacked a primer. The samples using L3 ligation reaction products and both pS-T1 and pS-T2 primers that were treated with SDS showed greater retention of product in the wells, consistent with denaturation of secondary structure in the RCA products.

Example 2

Simulation of Detection of Methylation Using Conversion of C to U by Bisulfite Treatment with a Linear Pair-locked Molecule

Determination of the sequence and 5-methylcytosine positions of a hypothetical duplex DNA fragment using conversion of C residues to U residues by bisulfite treatment is simulated as follows. The general scheme of this Example is illustrated in FIG. **12**. The sequence of the DNA is shown below.

DNA Sample (Methylated C Marked as ^{m}C)

```
5'-AGATGTGGA[m]CGGGGTGGG[m]CGGAGGTGGGTTGGGGC-3'  (SEQ ID NO: 5)
   |||||||||/ ||||||||/ ||||||||||||||||
3'-TCTACACCTG[m]CCCCACCCG[m]CCTCCACCCAACCCCG-5'  (SEQ ID NO: 6)
```

The two strands are connected by ligation to a linker sequence (represented as "nnnn") to give the following product. The linker sequence is suitable for use as a sequencing primer.

```
                                                (SEQ ID NO: 7)
3'-TCTACACCTG[m]CCCCACCCG[m]CCTCCACCCAACCCCGnnnnCGGGGT

TGGGTGGAGG[m]CGGGTGGGG[m]CAGGTGTAGA-5'
```

Additionally, a linear flap of known sequence (not shown) is attached to each end of the molecule of SEQ ID NO:7. The flap at the 3' end is suitable for primer binding for sequencing or replication. The complement of the flap at the 5' end is suitable for primer binding for sequencing or replication.

The product is treated with sodium bisulfite, resulting in the conversion of cytosine (but not 5-methylcytosine) residues to uracil, giving the following product. The newly formed uracil residues are bolded and marked with asterisks above the bases.

```
                                                (SEQ ID NO: 8)
    *  * **    *** ***   * ** ***  ****     *
3'-TUTAUAUUTG[m]CUUUAUUUG[m]CUTUUAUUUAAUUUUGnnnnUGGGGT

TGGGTGGAGG[m]CGGGTGGGG[m]CAGGTGTAGA-5'
```

A complementary strand (labeled SEQ ID NO: 9 below) is synthesized via DNA replication involving annealing of a primer to the flap added to the 3' end.

```
3'-TUTAUAUUTGCUUUAUUUGCUTUUAUUUAAUUUUGnnnnUGGGGTTGGGTGGAGGCGGGTGGGGCAGGTGTAGA-5'  (SEQ ID NO: 8)
   ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
5'-AAATATAAACGAAATAAACGAAAATAAATTAAAACnnnnACCCCAACCCACCTCCGCCCACCCCGTCCACATCT-3'  (SEQ ID NO: 9)
```

The above duplex is sequenced in both directions; sequencing intermediates are shown below. The nascent strand, whose sequence is being obtained, is SEQ ID NO: 10 in reaction a and SEQ ID NO: 11 in reaction b.

Thus, the reads predicted to be obtained from these reactions contain the following sequences.

```
                                                (SEQ ID NO: 10)
a: 5'-AGATGTGGACGGGGTGGGCGGAGGTGGGTTGGGGTnnnn-3'
```

```
                                                (SEQ ID NO: 11)
b: 5'-AAATATAAACGAAATAAACGAAAATAAATTAAAACnnnn-3'
```

The sequence of the original sample, including cytosine methylation status, is determined by applying the following rules, summarized in Table 5. The forward strand of the original sequence is the strand with the same sense as the two reads.

At positions where read a and read b both have A, the forward strand of the original sequence also has A, and the reverse strand has T. At positions where read a and read b both have T, the forward strand of the original sequence also has T, and the reverse strand has A.

When read a and read b both have C, then the forward strand of the original sequence has ^{m}C, and the reverse strand has G. When read a and read b both have G, then the forward strand of the original sequence has G, and the reverse strand has ^{m}C.

When one read has G at a position where the other read has A, the forward strand of the original sequence has G, and the reverse strand has C.

When one read has T at a position where the other read has C, the forward strand of the original sequence has C, and the reverse strand has G.

Reads a and b are matched to column 1 and 2 in Table 5 according to which read contains G and T residues at the positions where the reads differ; in this example, read a corresponds to column 1.

```
Sequencing reaction a

5'-AAATATAAACGAAATAAACGAAAATAAATTAAAACnnnnACCCCAACCCACCTCCGCCCACCCCGTCCACATCT-3'  (SEQ ID NO: 9)
                                  ||||||||||||||||||||||||||||||||||||||||
                              3'-nnnnTGGGGTTGGGTGGAGGCGGGTGGGGCAGGTGTAGA-5'  (SEQ ID NO: 10)
```

```
Sequencing reaction b

5'-AAATATAAACGAAATAAACGAAAATAAATTAAAACnnnn-3'                                  (SEQ ID NO: 11)
   |||||||||||||||||||||||||||||||||||||||
3'-TUTAUAUUTGCUUUAUUUGCUTUUAUUUAAUUUUGnnnnUGGGGTTGGGTGGAGGCGGGTGGGGCAGGTGTAGA-5'  (SEQ ID NO: 8)
```

TABLE 5

Bisulfite treatment methylation status determination rules

Sequencing reads		Original sequence	
		Forward strand	Reverse Strand
1	2	(5' => 3')	(3' => 5')
A	A	A	T
T	T	T	A
C	C	C methylated	G
G	G	G	C methylated
G	A	G	C
T	C	C	G

Application of the above rules to SEQ ID NOs: 10 and 11 results in recovery (after removal of the linker sequence nnnn) of the original sequences, i.e., SEQ ID NOs: 5 and 6. An alignment of reads a and b with the forward strand of the original sequence is shown in FIG. **10**A.

Example 3

Simulation of Detection of Methylation Using Conversion of mC to T by Photochemical Transition with a Linear Pair-locked Molecule

Determination of the sequence and 5-methylcytosine positions of a hypothetical duplex DNA fragment using conversion of ^{m}C to T by photochemical transition is simulated as follows. The general scheme of this Example is shown in FIG. **13**. The sequence of the DNA is shown below.

DNA Sample (Methylated C Marked as ^{m}C)

```
5'-AGATGTGGA(m)CGGGGTGGG(m)CGGAGGTGGGTTGGGGC-3'   (SEQ ID NO: 5)
   |||||||||/ ||||||||/ ||||||||||||||||
3'-TCTACACCTG(m)CCCCACCCG(m)CCTCCACCCAACCCCG-5'   (SEQ ID NO: 6)
```

The two strands are connected by ligation to a linker sequence (represented as "nnnn") to give the following product. The linker sequence is suitable for use as a sequencing primer. Linear flaps (not shown) are also attached to the 3' and 5' ends of this molecule.

```
                                                    (SEQ ID NO: 7)
3'-TCTACACCTG(m)CCCCACCCG(m)CCTCCACCCAACCCCGnnnnCGGGGT

TGGGTGGAGG(m)CGGGTGGGG(m)CAGGTGTAGA-5'
```

The product is treated with light so as to photochemically convert 5-methylcytosine (but not cytosine) residues to thymine, giving the following product. The newly formed thymine residues are bolded and marked with asterisks above or below the bases.

```
                                                    (SEQ ID NO: 12)
                  *         *
3'-TCTACACCTGTCCCACCCGTCTCCACCCAACCCCGnnnnCGGGGTTG

          *
GGTGGAGGTGGGTGGGGTAGGTGTAGA-5'
                    *
```

A complementary strand (labeled SEQ ID NO: 13 below) is synthesized via DNA replication using a primer that binds to the flap attached to the 3' end of the molecule.

```
3'-TCTACACCTGTCCCACCCGTCTCCACCCAACCCCGnnnnCGGGGTTGGGTGGAGGTGGGTGGGGTAGGTGTAGA-5' (SEQ ID NO: 12)
   ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
5'-AGATGTGGACAGGGTGGGCAGAGGTGGGTTGGGGCnnnnGCCCCAACCCACCTCCACCCACCCCATCCACATCT-3' (SEQ ID NO: 13)
```

The above duplex is sequenced in both directions as in Example 2 above, obtaining the following reads.

```
                                                    (SEQ ID NO: 14)
Read a: 5'-AGATGTGGATGGGGTGGGTGGAGGTGGGTTGGGGC-3'

                                                    (SEQ ID NO: 15)
Read b: 5'-AGATGTGGACAGGGTGGGCAGAGGTGGGTTGGGGC-3'
```

The sequence of the original sample, including cytosine methylation status, is determined by applying the following rules, summarized in Table 6. The forward strand of the original sequence is the strand with the same sense as the two reads.

At positions where read a and read b both have A, the forward strand of the original sequence also has A, and the reverse strand has T. At positions where read a and read b both have T, the forward strand of the original sequence also has T, and the reverse strand has A.

When read a and read b both have C, then the forward strand of the original sequence has C, and the reverse strand has G. When read a and read b both have G, then the forward strand of the original sequence has G, and the reverse strand has C.

When one read has G at a position where the other read has A, the forward strand of the original sequence has G, and the reverse strand has ^{m}C.

When one read has T at a position where the other read has C, the forward strand of the original sequence has ^{m}C, and the reverse strand has G.

Reads a and b are matched to column 1 and 2 in Table 6 according to which read contains G and T residues at the positions where the reads differ; in this example, read a corresponds to column 1.

TABLE 6

Photochemical transition methylation status determination rules

Sequencing reads		Original sequence	
		Forward strand	Reverse strand
1	2	(5' => 3')	(3' => 5')
A	A	A	T
T	T	T	A
C	C	C	G
G	G	G	C
G	A	G	C methylated
T	C	C methylated	G

Application of the above rules to SEQ ID NOs: 14 and 15 results in recovery (after removal of the linker sequence nnnn) of the original sequences, i.e., SEQ ID NOs: 5 and 6. An alignment of reads a and b with the forward strand of the original sequence is shown in FIG. **10**B.

```
a =>
                                                (SEQ ID NO: 14)
5'-AGATGTGGATGGGGTGGGTGGAGGTGGGTTGGGGC-3'

b =>
                                                (SEQ ID NO: 15)
5'-AGATGTGGACAGGGTGGGCAGAGGTGGGTTGGGGC-3'

r =>
                                                (SEQ ID NO: 5)
5'-AGATGTGGA'M'CGGGGTGGG'M'CGGAGGTGGGTTGGGGC-3' (r_a)

                                                (SEQ ID NO: 6)
3'-TCTACACCTG'M'CCCCACCCG'M'CCTCCACCCAACCCCG-5' (r_b)
```

Example 4

Comparison of the Accuracy of Simulated Single Read and Multiple Read Sequencing

The sequence of an assembled *Escherichia coli* genome, GenBank accession No. U00096, length 4639675 bp, was downloaded from GenBank. Randomly selected fragments with lengths ranging from 500 bp to 2000 bp were extracted from this sequence. These fragments were designated master sequences.

Five subsequences were generated from the master sequences by computationally introducing deletion and misreading errors at defined rates, as shown in Table 7.

The five subsequences, containing errors, were subjected to a multiple sequence comparison analysis using the CLUSTALW algorithm (default settings). The results of the CLUSTALW analysis were used as input for the program "cons" of the EMBOSS package in order to obtain a consensus sequence. The program "cons" is described in Rice et al., *Trends Genet* 16, 276-277 (2000), and Mullan et al., *Brief Bioinform* 3, 92-94 (2002).

The first subsequence and the consensus sequence were each compared to the master sequence, and the frequencies of gaps and misreads were tabulated; see Table 7. The results demonstrated that forming a consensus sequence using multiple reads reduced the frequency of misreads and gaps at each of the various error rates that was tested. For each set of deletion and misreading error rates, a single simulated read and a consensus sequence determined from 5 simulated reads were aligned against the master sequence. The number and percentage of misread and gapped positions were determined as a fraction of the positions in the alignment.

TABLE 7

Accuracy of consensus sequences determined from 5 simulated reads compared to individual reads at varying error rates

Rate of Introduced Errors	Length of Master (nt)	Single vs. Master		Consensus vs. Master	
		Misreads	Gaps	Misreads	Gaps
5% Deletion	816	53/816	47/816	8/817	5/817
1% Misreading		(6.5%)	(5.8%)	(1.0%)	(0.6%)
5% Deletion	1,565	90/1565	74/1565	9/1565	4/1565
2% Misreading		(5.8%)	(4.7%)	(1.0%)	(0.3%)
1% Deletion	1,589	401/1602	41/1602	90/1593	5/1593
30% Misreading		(25.0%)	(2.6%)	(5.6%)	(0.3%)
1% Deletion	760	182/76	11/76	47/761	1/861
30% Misreading		(23.8%)	(1.4%)	(6.2%)	(0.1%)

Example 5

Simulation of Determination of Sequence Using a cPLM

A double stranded nucleic acid sample is provided as in Example 2. The forward and reverse strands of the sample are locked together by ligation of an insert that forms a hairpin to each end of the molecule as shown in the cPLM construction step of FIG. **14** to form a circular pair-locked molecule. A single molecule sequencing by synthesis reaction is performed using a primer that binds to one of the inserts. Sequence data is obtained that comprises at least one sequence of the forward strand of the sample and at least one sequence of the reverse strand of the sample. The sequence data is analyzed by comparing the sequences of the forward and reverse strands of the circular pair-locked molecule to determine the sequence of the nucleic acid sample according to Table 8.

TABLE 8

cPLM sequence determination rules

Acquired sequence		Original sequence	
Templated by forward strand	Templated by reverse strand	Forward strand (5' => 3')	Reverse strand (3' => 5')
A	T	A	T
T	A	T	A
C	G	C	G
G	C	G	C

Note: in Table 8 and Tables 9-11 below, the acquired sequence templated by the forward strand corresponds to the upper line of sequencing data (i.e., the sequence shown beneath the arrow labeled "Sequencing" and above the arrow labeled "Sequence analysis") in FIGS. **14-17**, respectively. Similarly, the acquired sequence templated by the reverse strand corresponds to the lower line of sequencing data in FIGS. **14-17**, respectively.

Example 6

Simulation of Detection of Methylation Using Conversion of C to U by Bisulfite Treatment with a Circular Pair-locked Molecule

The general scheme of this Example is shown in FIG. **15**. A double stranded nucleic acid sample comprising at least one 5-methylcytosine is provided as in Example 2. A circular pair-locked molecule is formed as in Example 5. Bisulfite

conversion is performed as in Example 2. Sequence data is obtained as in Example 5. The sequence data is analyzed by comparing the sequences of the forward and reverse strands of the circular pair-locked molecule to determine the sequence of the nucleic acid sample and the position of the at least one 5-methylcytosine according to the rules in Table 9.

TABLE 9

cPLM/bisulfite treatment sequence determination rules

Acquired sequence		Original sequence	
Templated by reverse strand	Templated by forward strand	Forward strand (5' => 3')	Reverse strand (3' => 5')
A	T	A	T
T	A	T	A
C	A	C	G
A	C	G	C
C	G	G	C methylated
G	C	C methylated	G

Example 7

Simulation of Detection of Methylation Using Conversion of mC to T by Photochemical Transition with a Circular Pair-locked Molecule

The general scheme of this Example is shown in FIG. **16**. A double stranded nucleic acid sample comprising at least one 5-methylcytosine is provided as in Example 3. A circular pair-locked molecule is formed as in Example 5. Photochemical transition is performed as in Example 3. Sequence data is obtained as in Example 5. The sequence data is analyzed by comparing the sequences of the forward and reverse strands of the circular pair-locked molecule to determine the sequence of the nucleic acid sample and the position of the at least one 5-methylcytosine according to the rules in Table 10.

TABLE 10

cPLM/photochemical transition sequence determination rules

Acquired sequence		Original sequence	
Templated by reverse strand	Templated by forward strand	Forward strand (5' => 3')	Reverse strand (3' => 5')
A	T	A	T
T	A	T	A
C	G	C	G
G	C	G	C
C	A	G	C methylated
A	C	C methylated	G

Example 8

Simulation of Detection of 5-bromouracil Using a Circular Pair-locked Molecule

The general scheme of this Example is shown in FIG. **17**. A double stranded nucleic acid sample comprising at least one 5-bromouracil is provided. A circular pair-locked molecule is formed as in Example 5. Sequence data is obtained as in Example 5. The sequence data is analyzed by comparing the sequences of the forward and reverse strands of the circular pair-locked molecule to determine the sequence of the nucleic acid sample and the position of the at least one 5-bromouracil according to the rules in Table 11.

TABLE 11

cPLM/5-bromouracil sequence determination rules

Acquired sequence		Original sequence	
Templated by reverse strand	Templated by forward strand	Forward strand (5' => 3')	Reverse strand (3' => 5')
A	T	A	T
T	A	T	A
C	G	C	G
G	C	G	C
G	T	A	5-bromouracil
T	G	5-bromouracil	A

The specification is most thoroughly understood in light of the teachings of the references cited within the specification. The embodiments within the specification provide an illustration of embodiments of the invention and should not be construed to limit the scope of the invention. The skilled artisan readily recognizes that many other embodiments are encompassed by the invention. All publications and patents cited in this disclosure are incorporated by reference in their entirety. To the extent the material incorporated by reference contradicts or is inconsistent with this specification, the specification will supersede any such material. The citation of any references herein is not an admission that such references are prior art to the present invention.

Unless otherwise indicated, all numbers expressing quantities of ingredients, reaction conditions, and so forth used in the specification, including claims, are to be understood as being modified in all instances by the term "about." Accordingly, unless otherwise indicated to the contrary, the numerical parameters are approximations and may vary depending upon the desired properties sought to be obtained by the present invention. At the very least, and not as an attempt to limit the application of the doctrine of equivalents to the scope of the claims, each numerical parameter should be construed in light of the number of significant digits and ordinary rounding approaches. The recitation of series of numbers with differing amounts of significant digits in the specification is not to be construed as implying that numbers with fewer significant digits given have the same precision as numbers with more significant digits given.

The use of the word "a" or "an" when used in conjunction with the term "comprising" in the claims and/or the specification may mean "one," but it is also consistent with the meaning of "one or more," "at least one," and "one or more than one." The use of the term "or" in the claims is used to mean "and/or" unless explicitly indicated to refer to alternatives only or the alternatives are mutually exclusive, although the disclosure supports a definition that refers to only alternatives and "and/or."

Unless otherwise indicated, the term "at least" preceding a series of elements is to be understood to refer to every element in the series. Those skilled in the art will recognize, or be able to ascertain using no more than routine experimentation, many equivalents to the specific embodiments of the invention described herein. Such equivalents are intended to be encompassed by the following claims.

Unless defined otherwise, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. Although any methods and materials similar or equivalent to those described herein can be used in the practice or testing of the present invention, the preferred methods and materials are now described.

The publications discussed herein are provided solely for their disclosure prior to the filing date of the present application. Nothing herein is to be construed as an admission that the present invention is not entitled to antedate such publication by virtue of prior invention. Further, the dates of publication provided may be different from the actual publication dates which may need to be independently confirmed.

Other embodiments of the invention will be apparent to those skilled in the art from consideration of the specification and practice of the invention disclosed herein. It is intended that the specification and examples be considered as exemplary only, with a true scope and spirit of the invention being indicated by the following claims.

SEQUENCE LISTING

<160> NUMBER OF SEQ ID NOS: 19

<210> SEQ ID NO 1
<211> LENGTH: 79
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic oligonucleotide

<400> SEQUENCE: 1

cgacttatgc atttggtatc tgcgctctgc atatttaaat ggaaggagat agttaaggat 60

aagggcagag cgcagatac 79

<210> SEQ ID NO 2
<211> LENGTH: 80
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic oligonucleotide

<400> SEQUENCE: 2

caaatgcata agtcgtgtct taccgggttg atagcggccg ctcggagaaa agaagttgga 60

tgatgcaacc cggtaagaca 80

<210> SEQ ID NO 3
<211> LENGTH: 22
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic oligonucleotide

<400> SEQUENCE: 3

ccttatcctt aactatctcc tt 22

<210> SEQ ID NO 4
<211> LENGTH: 21
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic oligonucleotide

<400> SEQUENCE: 4

tagcggccgc tcggagaaaa g 21

<210> SEQ ID NO 5
<211> LENGTH: 35
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic oligonucleotide
<220> FEATURE:
<221> NAME/KEY: modified_base
<222> LOCATION: (10)..(10)
<223> OTHER INFORMATION: m5c
<220> FEATURE:
<221> NAME/KEY: modified_base
<222> LOCATION: (19)..(19)
<223> OTHER INFORMATION: m5c

-continued

<400> SEQUENCE: 5

agatgtggac ggggtgggcg gaggtgggtt ggggc 35

<210> SEQ ID NO 6
<211> LENGTH: 35
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic oligonucleotide
<220> FEATURE:
<221> NAME/KEY: modified_base
<222> LOCATION: (15)..(15)
<223> OTHER INFORMATION: m5c
<220> FEATURE:
<221> NAME/KEY: modified_base
<222> LOCATION: (24)..(24)
<223> OTHER INFORMATION: m5c

<400> SEQUENCE: 6

gcccaaccca cctccgccca ccccgtccaa catct 35

<210> SEQ ID NO 7
<211> LENGTH: 74
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic oligonucleotide
<220> FEATURE:
<221> NAME/KEY: modified_base
<222> LOCATION: (10)..(10)
<223> OTHER INFORMATION: m5c
<220> FEATURE:
<221> NAME/KEY: modified_base
<222> LOCATION: (19)..(19)
<223> OTHER INFORMATION: m5c
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: (36)..(39)
<223> OTHER INFORMATION: n is a, c, g, or t
<220> FEATURE:
<221> NAME/KEY: modified_base
<222> LOCATION: (55)..(55)
<223> OTHER INFORMATION: m5c
<220> FEATURE:
<221> NAME/KEY: modified_base
<222> LOCATION: (64)..(64)
<223> OTHER INFORMATION: m5c

<400> SEQUENCE: 7

agatgtggac ggggtgggcg gaggtgggtt ggggcnnnng ccccaaccca cctccgccca 60

ccccgtccac atct 74

<210> SEQ ID NO 8
<211> LENGTH: 74
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic oligonucleotide
<220> FEATURE:
<221> NAME/KEY: modified_base
<222> LOCATION: (10)..(10)
<223> OTHER INFORMATION: m5c
<220> FEATURE:
<221> NAME/KEY: modified_base
<222> LOCATION: (19)..(19)
<223> OTHER INFORMATION: m5c
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: (36)..(39)
<223> OTHER INFORMATION: n is a, c, g, t or u
<220> FEATURE:
<221> NAME/KEY: modified_base

<222> LOCATION: (55)..(55)
<223> OTHER INFORMATION: m5c
<220> FEATURE:
<221> NAME/KEY: modified_base
<222> LOCATION: (64)..(64)
<223> OTHER INFORMATION: m5c

<400> SEQUENCE: 8

agatgtggac ggggtgggcg gaggtgggtt ggggunnnng uuuuaauuua uutucguuua 60

uuucgtuuau atut 74

<210> SEQ ID NO 9
<211> LENGTH: 74
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic oligonucleotide
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: (36)..(39)
<223> OTHER INFORMATION: n is a, c, g, or t

<400> SEQUENCE: 9

aaatataaac gaaataaacg aaaataaatt aaaacnnnna ccccaaccca cctccgccca 60

ccccgtccac atct 74

<210> SEQ ID NO 10
<211> LENGTH: 39
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic oligonucleotide
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: (36)..(39)
<223> OTHER INFORMATION: n is a, c, g, or t

<400> SEQUENCE: 10

agatgtggac ggggtgggcg gaggtgggtt ggggtnnnn 39

<210> SEQ ID NO 11
<211> LENGTH: 39
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic oligonucleotide
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: (36)..(39)
<223> OTHER INFORMATION: n is a, c, g, or t

<400> SEQUENCE: 11

aaatataaac gaaataaacg aaaataaatt aaaacnnnn 39

<210> SEQ ID NO 12
<211> LENGTH: 74
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic oligonucleotide
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: (36)..(39)
<223> OTHER INFORMATION: n is a, c, g, or t

<400> SEQUENCE: 12

agatgtggat ggggtgggtg gaggtgggtt ggggcnnnng ccccaaccca cctctgccca 60

ccctgtccac atct 74

<210> SEQ ID NO 13
<211> LENGTH: 74
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic oligonucleotide
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: (36)..(39)
<223> OTHER INFORMATION: n is a, c, g, or t

<400> SEQUENCE: 13

agatgtggac agggtgggca gaggtgggtt ggggcnnnng ccccaaccca cctccaccca 60

ccccatccac atct 74

<210> SEQ ID NO 14
<211> LENGTH: 35
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic oligonucleotide

<400> SEQUENCE: 14

agatgtggat ggggtgggtg gaggtgggtt ggggc 35

<210> SEQ ID NO 15
<211> LENGTH: 35
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic oligonucleotide

<400> SEQUENCE: 15

agatgtggac agggtgggca gaggtgggtt ggggc 35

<210> SEQ ID NO 16
<211> LENGTH: 35
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic oligonucleotide
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: (15)..(15)
<223> OTHER INFORMATION: n is 5-bromouracil
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: (30)..(30)
<223> OTHER INFORMATION: n is 5-bromouracil

<400> SEQUENCE: 16

agatgtggac ggggngggcg gaggtgggtn ggggc 35

<210> SEQ ID NO 17
<211> LENGTH: 35
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic oligonucleotide
<220> FEATURE:
<221> NAME/KEY: misc_feature
<222> LOCATION: (27)..(27)
<223> OTHER INFORMATION: n is 5-bromouracil

<400> SEQUENCE: 17

gccccaaccc acctccgccc accccgncca catct 35

-continued

<210> SEQ ID NO 18
<211> LENGTH: 35
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic oligonucleotide

<400> SEQUENCE: 18

gccccaaccc acctccgccc gccccgtcca catct 35

<210> SEQ ID NO 19
<211> LENGTH: 34
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic oligonucleotide

<400> SEQUENCE: 19

agatgtgggc ggggtgggcg gaggtgggtg gggc 34

What is claimed is:

1. A method of determining a sequence of a double-stranded nucleic acid sample and a position of at least one modified base in the sequence, comprising:

a. locking the forward and reverse strands together to form a circular pair-locked molecule;

b. obtaining sequence data of the circular pair-locked molecule via single molecule sequencing, wherein the sequence data comprises sequences of the forward and reverse strands of the circular pair-locked molecule;

c. determining the sequence of the double-stranded nucleic acid sample by comparing the sequences of the forward and reverse strands of the circular pair-locked molecule;

d. altering the base-pairing specificity of bases of a specific type in the circular pair-locked molecule to produce an altered circular pair-locked molecule;

e. obtaining the sequence data of the altered circular pair-locked molecule wherein the sequence data comprises sequences of the altered forward and reverse strands; and

f. determining the positions of modified bases in the sequence of the double-stranded nucleic acid sample by comparing the sequences of the altered forward and reverse strands.

2. The method of claim **1**, wherein the double-stranded nucleic acid sample is obtained as a primary isolate from a cellular, viral, or environmental source.

3. The method of claim **2**, wherein the primary isolate is maintained at or below 25° C. in conditions substantially free of divalent cations and nucleic acid modifying enzymes prior to step (a) of claim **1**.

4. The method of claim **1**, wherein the double-stranded nucleic acid sample is obtained from an in vitro reaction or from extracellular nucleic acid.

5. The method of claim **1**, wherein altering the base-pairing specificity of bases of a specific type in the circular pair-locked molecule comprises bisulfite treatment.

6. The method of claim **1**, wherein altering the base-pairing specificity of bases of a specific type in the circular pair-locked molecule comprises photochemical transition.

7. The method of claim **1**, wherein locking the forward and reverse strands together comprises joining two nucleic acid inserts, which may be identical or non-identical, to the double-stranded nucleic acid sample, one to each end.

8. The method of claim **7**, wherein the nucleic acid inserts have lengths ranging from 14 to 200 nucleotide residues.

9. The method of claim **7**, wherein the nucleic acid inserts have known sequences.

10. The method of claim **7**, wherein the nucleic acid inserts form hairpins with overhangs, and the nucleic acid sample has overhangs compatible with the overhangs of the nucleic acid inserts.

11. The method of claim **7**, wherein obtaining sequence data comprises annealing a primer complementary to at least part of at least one of the nucleic acid inserts to the template and extending the primer.

12. The method of claim **7**, wherein at least one of the nucleic acid inserts comprises a promoter, and obtaining sequence data comprises contacting the promoter with an RNA polymerase that recognizes the promoter followed by synthesizing a product nucleic acid molecule comprising ribonucleotide residues.

13. The method of claim **7**, wherein joining is achieved by ligation.

14. The method of claim **1**, wherein the double-stranded nucleic acid sample comprises a plurality of samples linked together.

15. The method of claim **14**, wherein the samples of said plurality are linked via intervening nucleic acid inserts.

16. The method of claim **15**, wherein locking the forward and reverse strands together comprises ligating a complex formed by contacting the overhangs of the nucleic acid inserts with the compatible overhangs of the nucleic acid sample.

17. The method of claim **1**, wherein the double-stranded nucleic acid sample is a genomic DNA fragment.

18. The method of claim **1**, wherein the double-stranded nucleic acid sample comprises at least one RNA strand.

19. The method of claim **1**, wherein said single molecule sequencing comprises sequencing by a method chosen from single molecule sequencing by synthesis, and ligation sequencing.

20. The method of claim **1**, wherein said single molecule sequencing comprises real-time single molecule sequencing by synthesis.

21. The method of claim **1**, wherein said single molecule sequencing comprises single molecule sequencing by synthesis by a method chosen from pyrosequencing, reversible terminator sequencing, and third-generation sequencing.

22. The method of claim 1, wherein said single molecule sequencing comprises nanopore sequencing.

23. The method of claim 1, wherein:

the forward and reverse strands of the circular pair-locked molecule are locked together by nucleic acid inserts;

the sequence data obtained in step (b) comprise at least two copies of the sequence of the circular pair-locked molecule, each copy comprising sequences of first and second insert-sample units;

the sequences of the first and second insert-sample units comprise insert sequences, which may be identical or non-identical, and oppositely oriented repeats of the sequence of the nucleic acid sample; and

the method further comprises:

g. calculating scores of the sequences of at least four inserts contained in the sequence data by comparing the sequences of the at least four inserts to the known sequences of the inserts;

h. accepting or rejecting at least four of the repeats of the sequence of the nucleic acid sample contained in the sequence data according to the scores of one or both of the sequences of the inserts immediately upstream and downstream of the sample sequences, subject to the condition that at least one sample sequence in each orientation is accepted;

i. compiling an accepted sequence set comprising the at least one sample sequence in each orientation accepted in step (g); and

j. determining the sequence of the nucleic acid sample using the accepted sequence set.

24. A method of determining a sequence of a double-stranded nucleic acid sample and a position of at least one modified base in the sequence, comprising:

a. locking the forward and reverse strands of the nucleic acid sample together to form a circular pair-locked molecule;

b. obtaining sequence data of the circular pair-locked molecule via single molecule sequencing, wherein sequence data comprises sequences of the forward and reverse strands of the circular pair-locked molecule; and

c. determining the sequence of the double stranded nucleic acid sample and the position of the at least one modified base in the sequence of the double stranded nucleic acid sample by comparing the sequences of the forward and reverse strands of the circular pair-locked molecule, wherein at least one modified base in the double-stranded nucleic sample is paired with a base having a base pairing specificity different from its preferred partner base.

25. The method of claim 24, wherein the double stranded nucleic acid sample comprises at least one modified base chosen from 5-bromouracil, uracil, 5,6-dihydrouracil, ribothymine, 7-methylguanine, hypoxanthine, and xanthine.

26. A method of determining a sequence of a double-stranded nucleic acid sample and a position of at least one modified base in the sequence, comprising:

a. locking the forward and reverse strands of the nucleic acid sample together to form a circular pair-locked molecule;

b. altering the base-pairing specificity of bases of a specific type in the circular pair-locked molecule;

c. obtaining sequence data of the circular pair-locked molecule via single molecule sequencing, wherein sequence data comprises sequences of the forward and reverse strands of the circular pair-locked molecule; and

d. determining the sequence of the double-stranded nucleic acid sample and the position of the at least one modified base in the sequence of the double-stranded nucleic acid sample by comparing the sequences of the forward and reverse strands of the circular pair-locked molecule.

27. A method of determining a sequence of a double-stranded nucleic acid sample and a position of at least one modified base in the sequence, comprising:

a. locking the forward and reverse strands together to form a circular pair-locked molecule;

b. obtaining sequence data of the circular pair-locked molecule via single molecule sequencing, wherein the sequence data comprises sequences of the forward and reverse strands of the circular pair-locked molecule;

c. determining the sequence of the double-stranded nucleic acid sample by comparing the sequences of the forward and reverse strands of the circular pair-locked molecule;

d. obtaining sequencing data of the circular pair-locked molecule via single molecule sequencing, wherein at least one nucleotide analog that discriminates between a base and its modified form is used to obtain sequence data comprising at least one position wherein the at least one differentially labeled nucleotide analog was incorporated; and

e. determining the positions of modified bases in the sequence of the double-stranded nucleic acid sample by comparing the sequences of the forward and reverse strands.

28. A method of determining a sequence of a double-stranded nucleic acid sample and a position of at least one modified base in the sequence, comprising:

a. locking the forward and reverse strands of the nucleic acid sample together to form a circular pair-locked molecule;

b. obtaining sequence data of the circular pair-locked molecule via single molecule sequencing, wherein at least one nucleotide analog that discriminates between a base and its modified form is used to obtain sequence data comprising at least one position wherein the at least one differentially labeled nucleotide analog was incorporated; and

c. determining the sequence of the double-stranded nucleic acid sample and the position of the at least one modified base in the sequence of the double-stranded nucleic acid sample by comparing the sequences of the forward and reverse strands of the circular pair-locked molecule.

* * * * *

UNITED STATES PATENT AND TRADEMARK OFFICE

CERTIFICATE OF CORRECTION

PATENT NO. : 8,486,630 B2
APPLICATION NO. : 12/613291
DATED : July 16, 2013
INVENTOR(S) : Chao-Chi Pan et al.

It is certified that error appears in the above-identified patent and that said Letters Patent is hereby corrected as shown below:

IN THE CLAIMS:

Claim 24, col. 47, line 47, "nucleic sample is paired" should read --nucleic acid sample is paired--.

Signed and Sealed this
Third Day of September, 2013

Teresa Stanek Rea
Acting Director of the United States Patent and Trademark Office

CERTIFICATE OF SERVICE

I hereby certify that on March 19, 2015, the foregoing BRIEF OF APPELLANT INDUSTRIAL TECHNOLOGY RESEARCH INSTITUTE was filed electronically using the CM/ECF system and served via the CM/ECF system on registered counsel.

/s/ *Donna Stockton*

CERTIFICATE OF COMPLIANCE

I certify that the foregoing BRIEF OF APPELLANT INDUSTRIAL TECHNOLOGY RESEARCH INSTITUTE contains 13,529 words as measured by the word processing software used to prepare this brief.

Dated: March 19, 2015 Respectfully submitted,

/s/ Adam M. Breier

David S. Forman
FINNEGAN, HENDERSON, FARABOW, GARRETT & DUNNER, LLP
901 New York Avenue, NW
Washington, DC 20005-4413
(202) 408-4000

Erik R. Puknys
Adam M. Breier
FINNEGAN, HENDERSON, FARABOW, GARRETT & DUNNER, LLP
3300 Hillview Avenue
Palo Alto, CA 94304-1203
(650) 849-6600

Attorneys for Appellant